Modeling Protein Expression and Protein Signaling Pathways

Donatello Telesca[*] Peter Muller[†] Steven Kornblau[‡]
Marc Suchard[**] Yuan Ji[††]

[*]UCLA, dtelesca@ucla.edu
[†]UT Austin
[‡]MDACC
[**]UCLA
[††]MDACC

Modeling Protein Expression and Protein Signaling Pathways

Donatello Telesca, Peter Muller, Steven Kornblau, Marc Suchard, and Yuan Ji

Abstract

High-throughput functional proteomic technologies provide a way to quantify the expression of proteins of interest. Statistical inference centers on identifying the activation state of proteins and their patterns of molecular interaction formalized as dependence structure. Inference on dependence structure is particularly important when proteins are selected because they are part of a common molecular pathway. In that case inference on dependence structure reveals properties of the underlying pathway. We propose a probability model that represents molecular interactions at the level of hidden binary latent variables that can be interpreted as indicators for active versus inactive states of the proteins. The proposed approach exploits available expert knowledge about the target pathway to define an informative prior on the hidden conditional dependence structure. An important feature of this prior is that it provides an instrument to explicitly anchor the model space to a set of interactions of interest, favoring a local search approach to model determination. We apply our model to reverse phase protein array data from a study on acute myeloid leukemia. Our inference identifies relevant sub-pathways in relation to the unfolding of the biological process under study.

Modeling Protein Expression and Protein Signaling Pathways

DONATELLO TELESCA[1], PETER MÜLLER[2],
STEVEN KORNBLAU[3], MARC A. SUCHARD[1,5], YUAN JI[4]

Author's Footnote

[1] UCLA School of Public Health, Department of Biostatistics.

[2] The University of Texas at Austin, Department of Mathematics.

[3] The University of Texas M.D. Anderson Cancer Center, Department of Stem Cell Transplantation.

[4] The University of Texas M.D. Anderson Cancer Center, Department of Biostatistics.

[5] UCLA School of Medicine, Department of Biomathematics and Human Genetics.

October 18, 2011

For Correspondence

Donatello Telesca

e-mail: donatello.telesca@gmail.com

Yuan Ji

e-mail: yuanji@mdanderson.org

Modeling Protein Expression and Protein Signaling Pathways

Abstract

High-throughput functional proteomic technologies provide a way to quantify the expression of proteins of interest. Statistical inference centers on identifying the activation state of proteins and their patterns of molecular interaction formalized as dependence structure. Inference on dependence structure is particularly important when proteins are selected because they are part of a common molecular pathway. In that case inference on dependence structure reveals properties of the underlying pathway. We propose a probability model that represents molecular interactions at the level of hidden binary latent variables that can be interpreted as indicators for active versus inactive states of the proteins. The proposed approach exploits available expert knowledge about the target pathway to define an informative prior on the hidden conditional dependence structure. An important feature of this prior is that it provides an instrument to explicitly anchor the model space to a set of interactions of interest, favoring a local search approach to model determination. We apply our model to reverse phase protein array data from a study on acute myeloid leukemia. Our inference identifies relevant sub-pathways in relation to the unfolding of the biological process under study.

Keywords: AML; Graphical Models; Mixture Models; POE; RJ-MCMC; RPPA.

1 Introduction

In this article we consider the statistical analysis of high throughput protein expression data. We focus on the identification of patterns of protein interactions capitalizing on both a molecular interrogation protocol called reverse phase protein array (RPPA) technology and prior biological pathway knowledge.

RPPAs (Tibes et al. 2006) represent a high-throughput proteomic technology that provides a quantification of the expression for specifically targeted proteins selected from molecular pathways. Unlike traditional microarrays in which thousands of gene probes are immobilized on glass or fabric slides and are hybridized against samples from individuals, in RPPA one immobilizes individual samples on the slides and hybridizes them against a single antibody that recognizes only one protein. This is why the technology is named reverse phase. Figure 1 presents an image of RPPA, in which 40 individual patient samples containing whole cellular protein repertoires are printed in 40 batches. Each batch can be recognized as a square of four-by-four dots. For quantification purposes, each sample is diluted serially in eight steps so that the resulting strength of the dilutions are full strength, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, and 1/128 of the full strength. Duplicates are produced for each dilution, resulting in a total of 16 spots for each sample, shown as the four-by-four square on the slide. Each slide is probed with an antibody that represents one specific protein. The antibody can be detected by amplified fluorescent, colorimetric, or chemiluminescent assays. After each slide is hybridized against the antibody, signal intensities for all the samples on the slide are measured by scanning the slide using a specially designed scanner. Quantification of each sample is based on the intensities of the 16 spots. In our application we base our analysis on results obtained from the SuperCurve software (`http://bioinformatics.mdanderson.org/Software/OOMPA`). Currently, researchers are able to deposit over a thousand patient samples on one RPPA, making the technology particularly attractive for pathway analysis. In general the availability of protein-specific antibodies remains a limitation of RPPA, but this issue lies outside the scope of this manuscript.

Statistical research in the estimation of dependence patterns has mainly centered on Bayesian networks (Friedman 2004; Jansen et al. 2003; Pittman et al. 2004; Sebastiani and Ramoni 2005). Other recent attempts on network models include the work on finding differentially expressed gene-sets (often belonging to a genetic pathway) by Efron and Tibshirani (2006) and Newton et al. (2007). Recent extensions to this framework (Shojaie and Michaailidis, 2009 and 2010) show how the introduction of explicit prior information about gene-gene interactions improves power in the differential analysis of gene-sets. The innovative work by these authors still stops short of modeling specific interaction among the genes. In fact, in the setting of estimating gene-gene interactions, recent contributions have highlighted how well-established computational techniques often perform poorly in very high dimensional settings (Dobra et al. 2004, Scott and Carvalho 2008). In this regard, some progress has been reported when the analysis is carried out using convenient probabilistic schemes like Gaussian Markov random fields (Scott and Carvalho 2008; Wong et al. 2003; Jones et al. 2005), or functional transformation schemes (Telesca et al. 2009).

In the case of RPPA data, the curse of dimensionality is somewhat mitigated by the fact that the technology looks at a limited set (typically < 200) of proteins for expression and activation states, which represents a fraction of the estimated 4 million proteins and activation states thought to occur in the human proteome. Furthermore, in typical protocols, one is not dealing with the $n << p$ paradigm, as large sample sizes can easily be analyzed. On the other hand, when investigating dependence, some issues that are inherently associated with high throughput molecular interrogation persist. (1) The sampling distribution of proteomic abundance is usually non-Gaussian, rendering the convenient Gaussian graphical models framework inadequate as a model for protein interactions. (2) RPPA experiments often target specific biologic pathways; it is therefore important that the prior specification exploit available prior information on such targeted pathways.

We address issue (1) following the formalism introduced by Telesca et al. (2010), modeling the sampling distribution of protein abundance as a heavy tailed mixture distribution, and defining dependence at a hidden level, between binary indicators of inactive and active states. Addressing (2) defines a key feature of the proposed model, as our formulation introduces an informative prior, that assigns high prior probability mass to the model space surrounding the target pathway. Mukherjee and Speed (2008), for example, suggest a similar approach in the context of Gaussian DAGs. Our approach is also comparable to that described by Airoldi et al. (2007) in a data integration context. From a methodological perspective, we extend the work of Telesca et al. (2010) in two fundamental directions. First, we introduce a fully general dependence prior between hidden multivariate binary random quantities, and second, we propose a framework for Bayesian model determination that allows a local search for interactions, penalizing the model space as a function of a discrepancy measure between the target pathway and random departures from its graphical topology.

The rest of this article is structured as follows. We introduce the proposed analysis framework and probability model in § 2. Statistical inference based on MCMC sampling from the posterior distribution of interest is discussed in § 3. In § 4 we illustrate our approach through the analysis of an RPPA data set from a set of patients diagnosed with acute myeloid leukemia (AML). We conclude with a critical discussion of our contribution in § 5.

2 Dependent Mixture Model

We consider protein expression data in the form of a $G \times T$ matrix $\mathbf{Y} = [y_{gt}]$, with the generic element y_{gt} denoting the observed molecular expression for protein g in sample t, so that $g = 1, \ldots, G$ and $t = 1, \ldots T$. The sampling model for $\mathbf{Y}$ is defined conditionally on a latent variable e_{gt}, representing an indicator of class membership in a mixture density and is described in § 2.1. We argue that e_{gt} is likely to represent the biologically meaningful signal of protein activation better than the noisy row intensities y_{gt}. A key feature of the model is a graphical model $\mathcal{G}$ to parametrize a dependent prior on $\boldsymbol{e} = [e_{gt}]$. The graph $\mathcal{G}$ reflects the dependence structure between proteins. The strength of this dependence is indexed by an additional set of parameters $\boldsymbol{\beta}$. In summary the probability model is defined as:

$$p(\mathbf{Y}, \boldsymbol{e}, \boldsymbol{\beta}, \mathcal{G}) = \underbrace{p(\mathbf{Y} \mid \boldsymbol{e})}_{2.1} \; \underbrace{p(\boldsymbol{e} \mid \boldsymbol{\beta}, \mathcal{G})}_{2.2} \; \underbrace{p(\boldsymbol{\beta} \mid \mathcal{G})\, p(\mathcal{G})}_{2.3}. \tag{1}$$

Underbraced numbers indicate the subsections where each submodel is discussed below.

2.1 Sampling Model – $p(\mathbf{Y} \mid \boldsymbol{e})$

We assume that the sampling distribution of y_{gt} has heavy tails and follows a mixture of a normal distribution and two scale mixtures of uniform distributions. Let e_{gt} denote a latent trinary indicator, $e_{gt} \in \{-1, 0, 1\}$, indexing three possible distribution functions. Also, let $\tilde{y}_{gt} = y_{gt} - (\alpha_t + \mu_g)$, denote a normalized measure of relative abundance, corrected for a protein specific effect μ_g and a sample specific effect α_t. Using $f(\cdot)$ to generically denote a probability density function, we write the sampling distribution of $\tilde{y}_{gt}$ as:

$$f(\tilde{y}_{gt} \mid e_{gt}) = f_{-1}(\tilde{y}_{gt}) I(e_{gt} = -1) + f_0(\tilde{y}_{gt}) I(e_{gt} = 0) + f_1(\tilde{y}_{gt}) I(e_{gt} = 1), \tag{2}$$

where $f_{-1}(\cdot) = U(-k_g^-, 0)$, $f_0(\cdot) = N(0, \sigma_g^2)$, $f_1(\cdot) = U(0, k_g^+)$, $k_g^\pm > 0$ and $I(A)$ denotes an indicator of the event A. Such normal–uniform mixtures have been previously introduced by Parmigiani et al. (2002) and Dean and Raftery (2005) for the robust analysis of high throughput cDNA microarrays. Recently, Telesca et al. (2010) have extended the original formulation of Parmigiani et al. (2002), defining mixing proportions via a multivariate probit link.

Conditional on $\boldsymbol{e}_t = (e_{gt},\ g = 1, \ldots, G)$ the observed abundance measurements y_{gt} are exchangeable in (2). Later, in Section 2.2 we will exploit the trinary indicators e_{gt} to model conditional dependence between protein abundance measurements. For each protein g, the marginal distribution of $\mathbf{e}_g = (e_{g1}, ..., e_{gT})'$ is trinomial with T possible trials, support $\{-1, 0, 1\}$ and probabilities $\pi_g^- = p(e_{gt} = -1)$, $\pi_g^+ = p(e_{gt} = 1)$ and $\pi_g^0 = 1 - (\pi_g^+ + \pi_g^-)$.

In the foregoing formulation, σ_g^2 denotes the variance of the baseline distribution for protein g. The mixture distribution models overdispersion relative to $f_0(\cdot)$ through the tail parameters k_g^- and k_g^+. We will require that $\min(k_g^-, k_g^+) > 5\sigma_g$ to account for heavy tails in the proteomic distribution of abundance. More precisely, marginalizing over $\kappa_g^\pm$ the sampling model (2) is heavy tailed compared to the central normal distribution.

We propose the following conditionally conjugate hierarchical normal prior for the remaining parameters in (2): $\mu_g \sim N(m_\mu, \tau_\mu^2)$, $1/\sigma_g^2 \sim Ga(\gamma_\sigma, \lambda_\sigma)$, $\alpha_t \sim N(0, \tau_\alpha^2)$, $1/k_g^- \sim Ga(\gamma_k, \lambda_k)$ and finally $1/k_g^+ \sim Ga(\gamma_k, \lambda_k)$. Additionally, we will require $\sum_{t=1}^T \alpha_t = 0$, for likelihood identifiability.

2.2 Modeling Dependence – $p(\boldsymbol{e} \mid \boldsymbol{\beta}, \mathcal{G})$

We model dependence between proteins using the formalism of graphical models (Lauritzen 1996). A graph $\mathcal{G} = \{V, E\}$ is an algebraic structure, composed of a set of vertices V and a set of edges $E \subseteq V \times V$. When the set of vertices V represents a collection of random quantities, the edge structure E is often used to identify the full set of conditional independence relations between the components of V via, what is often called, the *global Markov property* associated with $\mathcal{G}$ (Besag 1974). Details, including whether edge set E is allowed to include directed edges, i.e., ordered pairs of nodes, or only undirected edges, i.e., unordered pairs, depend on the specific graphical model and are discussed below.

Regulatory relationships between proteins may include causal loops and dynamic reciprocal regulation (See Figure 2). This motivates the assumption that the set of conditional dependence relationships characterizing protein interactions is well represented by a class of graphical models known as *Reciprocal Graphs* (Koster 1996). These graphs allow for directed (e.g. $a \rightarrow b$) and undirected edges (e.g. $a - b$), loops (e.g. $a \rightarrow b$, $b \rightarrow c$, $c \rightarrow a$) and reciprocal relationships (e.g. $a \rightarrow b$, $b \rightarrow a$), where $a, b, c \in V$. Provided there are not directed edges between vertices of the same undirected path (e.g. $a - b$, $b - c$, $c \rightarrow a$), this class of models has a clear causal interpretation and equivalent Markov classes may be defined via graphical moralization (Lauritzen 1996, Koster 1996).

Our notation $\mathcal{G} = \{V, E\}$ defines a reciprocal graph. Each $\mathcal{G}$ corresponds to one and only one moral (undirected) graph $\mathcal{G}^m = \{V, E^m\}$, where E^m represents the set of undirected edges moralized from E. While we use a directed graph $\mathcal{G}$ to interpret the pathway diagram, in the absence of a time-course experiment, the data can only inform about an equilibrium distribution. For the development of the sampling model it therefore suffices to consider the implied conditional independence structure of the equilibrium distribution. We will represent this by G^m. We proceed in the following way. We start with a prior on G, this induces a prior on G^m. Conditional on G^m we define a sampling model for the observed data. Finally, posterior updating provides the desired posterior inference on G. In this process it is important to note that we only learn about the featurs of G that would change G^m. This process parallels inference in an ANOVA model that can only inform about identifiable contrasts. The prior on the directed graph can be thought of as prior regularization.

The mapping $\mathcal{G} \rightarrow \mathcal{G}^m$ is via the moralization procedure. Henceforth notation with superscript ^{m} will always refer to the moralized graph. The mapping is not one-to-one, as the same undirected graph may correspond to different directed structures, (Markov equivalence, Lauritzen 1996). In $\mathcal{G}^m$ standard Markov random field (mrf.) properties apply (Besag 1974). In particular, if we denote $ne(g)$ as the set of neighbors of protein g ($a \in ne(b)$ if and only if $\{a, b\} \in E^m$), then it is assumed that the abundance of protein g is conditionally independent of any other protein, given its neighbors $ne(g)$.

We specify a joint distribution over the latent indicators $\mathbf{e}_t = (e_{1t}, ..., e_{Gt})'$, $(t = 1, ..., T)$, that is consistent with the Markov structure spanned by a given reciprocal graph $\mathcal{G}$, through its moral representation $\mathcal{G}^m$. We first reduce the trinary e_{gt} to a binary indicator z_{gt}. We do so because in RPPA studies researchers usually focus on protein activations. We thus define an auxiliary indicator $z_{gt} = 2I\{e_{gt} = 1\} - 1 \in \{-1, 1\}$, which we will use to describe dependence across proteins.

Next we construct a joint probability model for $\boldsymbol{z}_t = (z_{gt},\ g = 1, \ldots, G)$. The reduction to the multivariate binary vector $\boldsymbol{z}_t$ greatly simplifies this modeling step. Following the arguments of Hammersley and Clifford (1971) and of Besag (1974), given G and the neighborhood structure encoded in a moral graph $\mathcal{G}^m$, the joint distribution $p(\mathbf{z}_t)$ can be defined in its full generality through the complete conditional probabilities $p(z_{gt} \mid \mathbf{z}_{(\backslash g,t)})$. More precisely, let $\exp[Q(\mathbf{z}_t)] \propto p(\mathbf{z}_t)$, then there exists an expansion of $Q(\boldsymbol{z}_t)$, unique on $\Omega = \{-1, 1\}^G$ and of the form:

$$Q(\boldsymbol{z}_t) = \sum_{1\le g\le G} z_{gt} H_g(z_{gt}) + \sum_{1\le g\le h\le G} z_{gt} z_{ht} H_{gh}(z_{gt}, z_{ht}) + \ldots + z_{1t} z_{2t} \cdots z_{Gt} H_{1,2,\ldots,G}(z_{1t}, z_{2t}, \ldots, z_{Gt}). \tag{3}$$

The Hammersley Clifford theorem (Besag 1974) postulates that while the H–functions may be chosen arbitrarily, for any set of labels $(1 \le g, h, ..., s \le G)$, $H_{g,h,...,s}$ may be non–null if and only if the set of proteins labeled $(g, h, ..., s)$ forms a *clique* in the moral graph $\mathcal{G}^m$. A clique is a subset of proteins in which each pair of proteins is connected by an edge.

The binary nature of z_{gt} allows us to, without loss of generality, replace any non–null H–function with a single arbitrary parameter (Besag 1974, Cox 1972) and we can therefore write:

$$Q(\boldsymbol{z}_t) = \sum \alpha_g z_{gt} + \sum \beta_{gh} z_{gt} z_{ht} + \cdots + \beta_{1,2,\ldots,G} z_{1t} z_{2t} \cdots z_{Gt}. \tag{4}$$

The foregoing formulation is a representation result for any multivariate distribution $p(\boldsymbol{z})$. Later, conditioning on a particular graphical model $\mathcal{G}$ will substantially reduce the number of terms appearing in (4). In fact, when considering $p(\boldsymbol{z}_t \mid \mathcal{G})$, the function $Q(\boldsymbol{z}_t \mid \mathcal{G})$ will only include sets of vertices forming cliques in the moralized graph $\mathcal{G}^m$. Finally, the definition of a joint distribution over the original trinary indicators $\mathbf{e}_t$ is completed with the conditional probability $p(e_{gt} = -1 \mid z_{gt} = -1) = \exp(\gamma_g)/\{1 + \exp(\gamma_g)\}$, $(g = 1, ..., G)$.

In summary, we have mapped $\mathcal{G}$ to $\mathcal{G}^m$, and then indexed all possible joint probability models on $\boldsymbol{e}_t$ that respect the conditional independence structure represented in $\mathcal{G}^m$ by (4).

2.3 Priors Over Interaction Parameters and Graphical Structures – $p(\boldsymbol{\beta} \mid \mathcal{G})\, p(\mathcal{G})$

We introduce an informative class of priors for the unknown graph $\mathcal{G}$. In words, our prior model is based on a pathway diagram that summarizes substantive prior information about the biochemical pathway of interest. First we interpret the pathway diagram as a known and fixed reciprocal graph $\mathcal{G}_0 = \{V, E_0\}$. Here we assume that the biochemical pathway is displayed as a set of nodes corresponding to proteins and edges between nodes. We assume that all edges are directed. In general, a reciprocal graph could also include undirected edges, subject only to the constraint that there be no directed edges between nodes of the same (undirected) pathway component (Koster, 1996). We do not make use of this feature in our implementation, as all edges in the original pathway are indeed directed.

Conceptually we write: $p(\mathcal{G}) \propto f\{d(\mathcal{G}_0, \mathcal{G})\}$, where $d(\cdot, \cdot)$ is a discrepancy measure and

$f' \leq 0$, so that graphs that deviate from $\mathcal{G}_0$ are assigned a lower prior probability. A similar approach was introduced by Mukherjee and Speed (2008) for Gaussian DAGs.

Let A^c denote the complement of set A and define $d(\mathcal{G}_0, \mathcal{G}) = |E^c \cap E_0| + \delta\, |E \cap E_0^c|$, $\delta > 1$. Here $|E|$ identifies the number of edges in a graph $\mathcal{G} = \{V, E\}$. This defines a discrepancy measure $d(\mathcal{G}_0, \mathcal{G})$ as the weighted sum of edges dropped from $\mathcal{G}_0$ and edges added to $\mathcal{G}_0$. If we assume for a moment $\delta = 1$, then $d(\mathcal{G}_0, \mathcal{G})$ reduces to the number of changed edges relative to $\mathcal{G}_0$. A weight $\delta > 1$ allows to include a notion of parsimony by using an increased penalty for adding edges compared to dropping edges. For a given $d(\cdot, \cdot)$, we assume that

$$p(\mathcal{G} \mid \varphi) \propto f\{d(\mathcal{G}_0, \mathcal{G})\} = \varphi^{d(\mathcal{G}_0, \mathcal{G})}, \tag{5}$$

where we assume $\varphi \sim \mathcal{B}(a_\varphi, b_\varphi)$. The prior (5), with $\delta = 1$, is equivalent to assuming exchangeable coin flips to determine each edge, with odds of matching edges in $\mathcal{G}_0$ equal to φ. Similar probability models have been shown to provide in several contexts automatic multiplicity correction in the posterior (Scott and Berger 2008). For any given φ, the expected density of a graph [1] $\mathcal{G}$ is computed as $\bar{d}\,(\mathcal{G} \mid \varphi) = \left(|E_0^c|\, \frac{\varphi^\delta}{1+\varphi^\delta} + \frac{|E_0|}{1+\varphi}\right) \frac{1}{G(G-1)}$, which is a decreasing function of δ. Furthermore, the distribution of the graph density $d(\mathcal{G} \mid \varphi)$ has mode at $\frac{|E_0|}{G(G-1)}$ and variance decreasing with δ. A key property of (5) is therefore the informative nature of the prior, maximized at the prior guess $\mathcal{G}_0$, leading to posterior simulation being a local exploration of the model space. This is in sharp contrast to alternative approaches that allow for global model exploration by using prior probability models that are far more diffuse over the model space. If desired, further structural restrictions are easily incorporated in this framework by introducing changes in $d(\mathcal{G}_0, \mathcal{G})$. For example, this function could be defined in ways that force the inclusion of a subset of edges, say E_0^*, or in ways that introduce discrete cut-offs to the overall model size.

[1] The density of a reciprocal graph $\mathcal{G}$ is defined as $d(\mathcal{G}) = \frac{|E|}{G(G-1)}$.

Our model is completed by defining normal priors over the interaction parameters $(\beta_{gh}, \beta_{ghk}, ...)$, $(\forall\ g < h < k \cdots)$, so that $\beta_u \mid \mathcal{G} \sim N(\eta_u, \sigma^2_\beta)$, for any generic index u of the parameter vector $\boldsymbol{\beta}$. The mean value η_u is used to represent an important feature of typical molecular pathway representations. Pathway diagrams usually include a distinction between stimulatory and inhibitory interactions. In Figure 2, for example, edges between proteins are labelled with an arrowed-tip $(i \rightarrow j)$ if interactions observed a priori are of a stimulatory nature and with a bullet-tip (—●) if prior interactions are inhibitory.

We formalize the definition of stimulatory and inhibitory relationships, following the framework introduced in § 2.2 and focusing on second order interactions parameters β_{gh}, $(g < h)$. More precisely, indexing protein activation with binary indicators $z_{gt} \in \{-1, 1\}$, we assume that protein g stimulates the activation of protein h if $p(z_{gt} = 1 \mid z_{ht} = 1) > p(z_{gt} = 1|z_{ht} = -1)$. Similarly, for inhibitory relationships we assume $p(z_{gt} = -1 \mid z_{ht} = 1) > p(z_{gt} = -1|z_{ht} = -1)$. It is easy to verify that these definitions motivate the choice of $\eta_u > 0$ for stimulatory and $\eta_u < 0$ for prior inhibitory interactions. In the absence of prior information, η_u is simply set to 0.

For reference, we summarize the complete model. Let $\boldsymbol{\lambda}_g = (\kappa^-_g, \kappa^+_g, s_g)'$; we have:

$$\begin{array}{ll}
\text{Sampling model:} & \tilde{y}_{gt} \mid e_{gt} = e, \boldsymbol{\lambda}_g \sim f_e(\tilde{y}_{gt}), \quad \tilde{y}_{gt} = y_{gt} - (\alpha_t + \mu_g) \\
\text{Autologistic:} & p(\boldsymbol{z}_t \mid \mathcal{G}, \boldsymbol{\beta}) \propto \exp\{\sum_g \alpha_g z_{gt} + \sum_{(g,h)\in E^m} \beta_{gh} z_{gt} z_{ht}\} \\
\text{Graphical model prior:} & p(\mathcal{G} \mid \varphi) \propto \varphi^{d(\mathcal{G}_0, \mathcal{G})} \\
\text{Hyperprior:} & \beta_{ij} \mid \mathcal{G} \sim N(\eta_{ij}, \sigma^2_\beta),\ (i, j) \in E^m
\end{array}$$

where the trinary indicator e_{gt} is an elaboration of the binary indicator z_{gt} for activation, with $\log \frac{p(e_{gt}=-1|z_{gt}=-1)}{p(e_{gt}=0|z_{gt}=-1)} = \gamma_g$ and $e_{gt} = 1$ when $z_{gt} = 1$.

3 Posterior Computation and Inference

3.1 Stochastic Search and MCMC Computation

We are interested in identifying patterns of molecular interactions, as informed by a prior pathway. The full model is determined by mixture membership indicators $\boldsymbol{e}$, a conditional dependence configuration summarized in a graph $\mathcal{G}$, MRF parameters $\boldsymbol{\beta}$ and a collection of parameters $\boldsymbol{\theta}$, which we assume contains all remaining random quantities with the exception of $\mathbf{Y}$. Our inference centers on the posterior distribution $p(\boldsymbol{e}, \mathcal{G}, \boldsymbol{\beta}, \boldsymbol{\theta} \,|\, Y)$, which fully summarizes the available evidence on protein abundance and protein expression profiles similarities.

The posterior probability model is not available in closed form. However, it is conceptually straightforward to define MCMC simulation from this target distribution. In particular, we could proceed as usual and draw the parameters of interest sequentially or in random order from the full conditional posterior distributions of $\boldsymbol{\theta}$, $\boldsymbol{e}$, $\boldsymbol{\beta}$ and $\mathcal{G}$. Updating $\boldsymbol{e}$ and $\boldsymbol{\theta}$ can be implemented via standard Gibbs sampling.

Unfortunately, updating $\boldsymbol{\beta}$ and $\mathcal{G}$ remains problematic (Green and Richardson 2002). The complete conditional posterior for $\boldsymbol{\beta}$ and $\mathcal{G}$ depend on $\boldsymbol{e}$ only indirectly through $\boldsymbol{z}$. Recall that we further summarized the trinary indicators $\boldsymbol{e}$ into binary indicators $\boldsymbol{z} \in \{-1, 1\}^G$ of protein activation. The conditional posterior of $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta}|Y, \boldsymbol{z}, \mathcal{G}, \boldsymbol{\theta}) = p(\boldsymbol{\beta}|\boldsymbol{z}, \mathcal{G}, \boldsymbol{\theta}) \propto p(\boldsymbol{z}|\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\theta})\; \pi(\boldsymbol{\beta}|\mathcal{G}, \boldsymbol{\theta}), \tag{6}$$

where

$$p(\boldsymbol{z}|\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\theta}) = \prod_{t=1}^{T} \frac{\exp\{Q(\boldsymbol{z}_t)\}}{g(\boldsymbol{\beta})}, \qquad g(\boldsymbol{\beta}) = \sum_{\boldsymbol{z}} \exp\{Q(\boldsymbol{z}_t)\}$$

and the sum $\sum_{\boldsymbol{z}}$ is over all 2^G possible realizations of $\boldsymbol{z}$.

The full conditional distribution of $\boldsymbol{\beta}$ is therefore defined in terms of a partition function

$g(\boldsymbol{\beta})$, which requires summation over all possible 2^G realizations of $\boldsymbol{z}_t$, for any $t = 1, ..., T$. This quantity can be evaluated efficiently only for a very small G. This is often not the case in RPPA studies, which leaves us with the problem of devising a feasible strategy to update $\boldsymbol{\beta}$. Similar considerations apply for updates involving $\mathcal{G}$, with the further complication that changes in $\mathcal{G}$ define changes in the dimensionality of the parameter vector $\boldsymbol{\beta}$.

This problem is well known in the literature on Markov random fields and several solutions have been proposed, based on ad hoc approximation to the partition function of interest (Green and Richardson 2002, Friel et al. 2009). In the following we propose a solution based on MCMC ingenuity and show that, exploiting a transition kernel, that updates mixture indicators $\boldsymbol{e}$ and dependence parameters $\boldsymbol{\beta}$ in a joint fashion is possible without calculation of $g(\boldsymbol{\beta})$.

3.2 Updating $\boldsymbol{\beta}$

To update $\boldsymbol{\beta}$ we consider a Metropolis-Hastings (MH) transition kernel that involves also changes in $\boldsymbol{z}$ and consequently $\boldsymbol{e}$. For any generic element of $\boldsymbol{\beta}$, say β_u ($u \subset \{1, \cdots, M\}$), we consider a proposal distribution $(\beta'_u, \boldsymbol{e}') \sim q(\beta'_u)\, p(\boldsymbol{e}'|\beta'_u, \boldsymbol{\beta}_{(\backslash u)}, \mathcal{G}, \boldsymbol{\theta})$ where $q(\cdot)$ is an arbitrary proposal density and possibly dependent on the previous values β_u, whereas $p(\boldsymbol{e}|\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\theta})$ is the conditional prior defined in § 2.2. In our implementation, for all $i > j = 1, ..., G$, we consider a random walk proposal $q(\beta'_{ij}) =_d N(\beta_{ij},\ s_\beta)$, where s_β is calibrated at burn-in to achieve acceptance rates between 30% and 70%.

Draws from $\boldsymbol{e}'|\boldsymbol{\beta}', \mathcal{G}, \boldsymbol{\theta}$ may then be obtained easily via a short Gibbs run, based on full conditionals $p(e_{gt}|\boldsymbol{e}_{\backslash g}, \boldsymbol{\beta}', \mathcal{G}, \boldsymbol{\theta})$. It is easy to verify that for the construction of a standard Metropolis-Hastings transition kernel the newly proposed values $(\boldsymbol{e}',\ \beta'_{ij})$ must be accepted with probability

$$\rho(\beta_u, \boldsymbol{e}, \beta_u', \boldsymbol{e}') = \min \left\{ \frac{p(Y \mid \boldsymbol{e}', \boldsymbol{\theta})}{p(Y \mid \boldsymbol{e}, \boldsymbol{\theta})} \frac{\pi(\beta_{ij}' \mid \mathcal{G}, \boldsymbol{\theta})}{\pi(\beta_{ij} \mid \mathcal{G}, \boldsymbol{\theta})} \frac{q(\beta_{ij})}{q(\beta_{ij}')};\ 1 \right\}. \tag{7}$$

The structure of the proposal distribution allows for the elimination of the partition function $g(\boldsymbol{\beta})$, at the cost of having to simulate $G \times T$ mixture indicators for each β_{ij}. Gains in efficiency could be further achieved considering block updates for the elements of $\boldsymbol{\beta}$. Our approach is similar to that of Møller et al. 2006.

In short, we replace the problem of evaluating the partition function with a problem of approximate prior simulation. The resulting MCMC provides an approximation to the desired posterior only to the extent to which the prior Gibbs simulation can be considered a draw from the prior. A systematic discussion of related MCMC schemes appears in Andriew and Roberts (2009), who also discuss several more elaborate variations of this strategy. Further precision may be achieved considering exact simulation from $\boldsymbol{e}|\boldsymbol{\beta}, \mathcal{G}, \boldsymbol{\theta}$ (Propp and Wilson 1996), but this would come at a higher computational cost. In fact, the asymptotic validity of the proposed computation and a favorable comparison to exact sampling methods have recently been discussed by Liang (2010). A comparative review of alternative computational strategies associated with similar probability models (Ising and Potts models) has been compiled by Zhou and Schmidler (2009).

3.3 Updating $\mathcal{G}$ via RJ-MCMC

Updating $\mathcal{G}$ involves changes in $\boldsymbol{\beta}$ and its dimensionality. We are therefore faced with two complications, one associated with the evaluation of the partition function $g(\boldsymbol{\beta})$, as explained above, and the other associated with the need to maintain detailed balance across dimensions. To solve this problem, we combine the reversible jumps algorithm of Green (1995) with the approach described in § 3.2. More precisely, we follow Giudici and Green (1999) and propose the following transitions:

1. Select an edge (i, j) at random. If $(i, j) \in E$ propose its elimination, otherwise if $(i, j) \notin E$ propose its birth. This corresponds to moving from the current graph $\mathcal{G}$ to a new graph $\mathcal{G}'$, with moral representation $\mathcal{G}^{m\prime}$.

2. Assume the move in step 1. involves a birth. Propose a set of, say k, new interaction parameters $\boldsymbol{\beta}_k' \sim q(\boldsymbol{\beta}_k')$, in correspondence to new undirected edges in the moral graph $\mathcal{G}^{m\prime}$. In our implementation, we consider a proposal density based on local logistic regressions. More precisely, if we propose the birth of an interaction parameter β_{ij} involving $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, we define: $a = \sum_t I\{z_{it} = -1, z_{jt} = -1\}$, $b = \sum_t I\{z_{it} = -1, z_{jt} = 1\}$, $c = \sum_t \{z_{it} = 1, z_{jt} = -1\}$ and $d = \sum_t I\{z_{it} = 1, z_{jt} = 1\}$. We calculate $\hat{\beta}_{ij} = log(\frac{ad}{bc})/2$ and $\delta_\beta = \sqrt{(1/a + 1/b + 1/c + 1/d)}$. Here $\hat{\beta}_{ij}$ estimates the log odds ratio for a logistic regression involving $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. The parameter δ_β is a standard large sample approximation to $SE(\hat{\beta}_{ij})$. The proposal density is then constructed as a Normal approximation to the conditional likelihood surface, so that $q(\beta_{ij}') =_d N(\hat{\beta}_{ij}, \delta_\beta^2)$.

3. Propose new values for the mixture indicators $\boldsymbol{e}' \sim p(\boldsymbol{e}'|\boldsymbol{\beta}', \mathcal{G}', \boldsymbol{\theta})$. Draws are made from the conditional prior defined in § 2.2.

4. Accept steps 1., 2. and 3. with probability $\rho_b(\mathcal{G}, \mathcal{G}')$:

$$\rho_b(\mathcal{G}, \mathcal{G}') = \min \left\{ \frac{p(Y \mid \boldsymbol{e}', \boldsymbol{\theta})}{p(Y \mid \boldsymbol{e}, \boldsymbol{\theta})} \frac{\pi(\mathcal{G}')}{\pi(\mathcal{G})} \frac{p(\beta_k')}{q(\boldsymbol{\beta}_k')};\ 1 \right\}.$$

The reverse move, involving the death of a random edge (i, j), is implemented by setting to 0 the elements of $\boldsymbol{\beta}$ disappearing from $\mathcal{G}^{m\prime}$ and it is accepted with probability $\rho_d(\mathcal{G}, \mathcal{G}') = 1/\rho_b(\mathcal{G}, \mathcal{G}')$.

The transition probability for $\mathcal{G}$ includes a proposal for elements of $\boldsymbol{\beta}$. One could therefore drop the transition probability described in §3.1.1, and still maintain irreducibility. However, without 3.1.1. the resulting MCMC demonstrates far slower mixing.

We conclude by summarizing the overall MCMC. We use notation $\mathbf{x}_{-i}$ to indicate the vector $\mathbf{x}$ without the element x_i and $[x \mid y, z]$ to generically indicate a transition probability that changes x and depends on the currently imputed values of y, z. Each iteration in the proposed posterior simulation algorithm involves the following transition sequence:

$$[\boldsymbol{\theta} \mid \boldsymbol{e}, \beta, \mathcal{G}^m, \mathbf{Y}],\ [e_{gt} \mid \boldsymbol{e}_{-gt}, \boldsymbol{\beta}, \mathcal{G}^m, \boldsymbol{\theta}, \mathbf{Y}],\ [\beta_{ij} \mid \boldsymbol{\beta}_{-ij}, \boldsymbol{e}, \boldsymbol{\theta}, \mathcal{G}^m],\ [\mathcal{G} \mid \boldsymbol{e}, \boldsymbol{\beta}, \boldsymbol{\theta}].$$

4 Simulated Data

We carried out a small simulation experiment to validate inference under the proposed model, to investigate the impact of prior choices, and to compare with alternative approaches. Details are reported in the supplementary materials to this paper. In brief, we simulated data under a model that differs from the proposed model and misspecified our priors on network structures. Figure 3 summarizes some aspects of this study. Panel (a) validates inference by comparing the inference on edge inclusion with the simulation truth. The figure shows the ROC curve for classifying proteins into activated and non-activated. Panel (b) shows ROC curves for selecting edges in the graph and compares inference under the proposed model, under GeneNet (Schäfer and Strimmer 2005), and under a model with exchangeable Bernoulli priors on all edges. Panel (c) shows similar ROC curves under alternative priors with the proposed model, comparing the proposed informative prior (5), and two non-informative priors.

5 Example: Analysis of RPPA Data

5.1 A Study in Acute Myeloid Leukemia

The current classification of AML uses the French-American-British system based on morphological features, along with flow cytometric analysis of surface marker expression, cytogenetics, and assessment of recurrent molecular abnormalities. These classification schemes have prognostic relevance but, with the exception of acute promyelocytic leukemia, they generally do not alter therapeutic recommendations (Mrozek et al. 2007). Furthermore, the predictive abilities of prognostic models based on current clinical and laboratory features are generally low, with less than half of the outcomes explained by these features. Researchers have recently started to investigate genomic and proteomic features for prognosticating AML. Proteins within signaling pathways that exhibit heterogeneous expression in AML are often prognostic, a characteristic that has been studied by researchers such as Kornblau et al. (2006) and Tanner et al. (2001). These authors find that distinct molecular abnormalities and patterns of pathway activation in leukemic cells collectively suggest potential targets for therapeutic intervention. Consequently, knowledge allowing rational evaluation of targeted therapies on an individualized basis in AML is sorely needed. As part of a knowledge-learning process, expression levels and activation of a single protein or a limited number of proteins have been studied by Kornblau et al. (1994), Kornblau et al. (2000), and Kreuter et al. (2006).

The natural next step is formal inference on different patterns of pathway activation and modification. This requires inference on the joint probability distribution for multiple proteins in a pathway. The proposed model and inference approach is built to facilitate such inference.

We use data from a large AML study based on RPPA. Specifically, we have probed protein samples from 531 newly diagnosed, primary refractory, and relapsed AML patients. The objective of this experiment was to investigate interactions of important protein markers related to AML. We selected 51 proteins in signal transduction, apoptosis, and cell cycle regulatory pathways (Figure 2) and studied their expression profiles in all 531 samples. An important feature of the AML data under study is that the number of samples ($n = 531$) is much greater than the number of proteins ($p = 51$). This feature facilitates principled model-based inference and model assessment and contrasts inference for RPPA data with inference for many other high throughput genomic platforms.

The targeted interactions are illustrated in Figure 2, where we present a comprehensive signaling network map for the selected proteins based on the three pathways. The signaling network map was developed to show signaling interactions based on published articles from PubMed searches as well as from the connections map in Signal Transduction Knowledge Environment (http://www.stke.org).

5.2 Data Analysis

Using a desktop scanner at an optical resolution of 1200 dpi, we scanned the hybridized RPPA slides and saved them as TIFF files. The protein expression intensity of each spot was measured with an automated software program, MicroVigene™ (VigeneTech, Inc. North Billerica, MA). The dilution series of the samples provided a dilution-concentration-expression curve, providing relative expression intensities that were read off in the linear part of the curve. We used these numbers for data preprocessing and calculation after standardization and topographical normalization (for details, refer to the SuperCurve software at `http://bioinformatics.mdanderson.org/Software/OOMPA`). Good pre-processing of data from any high throughput experiments in general, and RPPA data in particular, is critically important. We used SuperCurve as the best currently available model-based and principled implementation of inference to quantify RPPA data.

We estimate protein pathway structure by combining prior knowledge of protein interaction and RPPA measurements. Specifically, given the prior consensus pathway in Figure 2, we use the proposed probability model to update our prior knowledge, combining the qualitative information about the consensus pathway with quantitative information from RPPA data (Figure 1). We fit the model presented in § 2 to this set of 51 proteins. Reported inference is based on 500,000 MCMC samples (thinned by 10) after discarding 100,000 samples for burn-in.

In Figure 4 (left panel) we illustrate the fit of the proposed model to the AML protein expression data reported in Figure 1. Defining $\pi^+_{gt} = p(e_{gt} = 1|Y)$ and $\pi^-_{gt} = p(e_{gt} = -1|Y)$ we follow Parmigiani et al. (2002) and define a probability of expression scale $p^*_{gt} \in [-1, 1]$. This quantity represents a univariate posterior summary, designed to relate directly to the mixture indicator labels $e_{gt} \in \{-1, 0, 1\}$ and may be considered as the normalized scale for raw intensities y_{gt}. Here we show how the original abundance measurement is translated into a probability of expression on the new scale (p^*). Plotting the raw correlation coefficients versus the the simple correlation estimated in the p^* scale (Figure 4, right panel), we note that the two measures of association are clearly correlated, with a tendency of stronger agreement towards large absolute values of the correlation coefficients.

Our inference on molecular interaction is based on posterior edge inclusion probabilities, $P\{(i, j) \in E \ or \ (j, i) \in E \mid Y)\}$. These quantities are estimated directly from the RJ-MCMC output as the percentage of iterations where the edge is included in the current graphical structure $\mathcal{G}$. In Figure 5 we represent the posterior edge inclusion probabilities for all possible protein interactions, against expected posterior interactions $E(\beta_{ij} \mid \mathbf{Y})$. Solid diamonds denote edges originally included in the prior pathway $\mathcal{G}_0$. Our differential penalization scheme ($\delta = 5$) favors edges in the target pathway, but still allows for the model to explore interactions which are not included in the prior. Reporting a final pathway may be based on several criteria (Telesca et al. 2009). Here we consider the median model (Barbieri and Berger 2004, Scott and Berger 2008). That is, we select edges with posterior inclusion probability greater than 0.5.

The selected posterior network is reported in Figure 5. The edge thickness is proportional to the absolute value of the associated expected posterior interaction parameter $E(\beta_{ij} \mid \mathbf{Y})$. We distinguish between stimulatory and inhibitory relationships. Following the argument introduced in § 2.3, we base our inference on the sign of the posterior mean $E(\beta_{ij} \mid Y)$. For a generic edge index u, a positive posterior mean is interpreted as stimulation, whereas a negative sign is interpreted as inhibition. In the reported figure, we represent stimulation with arrows and inhibition with dotted arrowheads.

Our analysis identifies two main sub-pathways (Akt/mTOR and STATs) as groups of proteins exhibiting significant similarities in their over-abundance patterns across samples. The Akt/mTOR pathway is known to be a key player in cell growth, proliferation and survival. Recent literature identifies this pathway as a strong contributor to proliferation and drug resistance in AML (a comprehensive review is provided by Martelli et al. (2009)). The second active pathway involves the family of STAT proteins (STAT3 / STAT5). These molecules are known to act as important regulators in hematopoiesis (blood cells formation) and their up-regulation via β-catenin (BCAT) has been documented by several authors (Hao et al. 2006, Boeuf et al. 2001). In the same pathway we also recover key interactions between the STAT proteins and extracellular-signal regulated kinase (ERK) (Jain et al. 1998). The important role of Stat5 to AML was also recently noted by Kornblau et al. (2010).

Some of our results are surprising. For example, PTEN.p is inactivated when phosphorylated and therefore cannot inhibit AKT phosphorylation. The negative relationship between

PTEN.p and AKT.p308 seems therefore contrary to canonical expectations (Wan and Helman 2003). The PTEN-AKT pathway is however of great interest in oncology and this finding may prove useful in the search for possible feedback regulatory interactions or latent oncogenes (see for example the discussion in Palomero et al. (2008)). Likewise, one would expect the pAKTs to have an edge with pGSK3 instead of total AKT. On the other hand, the inferred diagram confirms evidence that, in active form, AKT phosphorylates a wide variety of downstream substrates. Also the diagram highlights MTOR as a main activator of the serine P70S6K, possibly contributing to tumor cell survival. Furthermore, the fact that MTOR signaling is confirmed to operate downstream, makes it a promising therapeutic target for AML patients (Chen et al. 2010).

These unexpected findings may simply represent misidentified relationships or they may reflect regulatory feedback within the system. At the same time, however, they may define the basis for genuinely novel discoveries. While this analysis is exploratory in nature, we believe it provides a principled hypothesis generation instrument for further experimental investigation. This instrument may prove particularly significant in a disease like AML, where standard chemotherapy is too often not fully effective and where there is great need for new therapeutic strategies.

6 Discussion

We have proposed a probabilistic framework for the analysis of RPPA data. We focus on assessing the probability of protein–protein interactions considering patterns of similarity characterized by hidden states of activation (as compared to inactive states).

Our probability model makes explicit use of prior information regarding evidence of protein interactions reported in the literature. This allows for the definition of a prior over graphical structures that explicitly anchors the model to known patterns of interaction, still allowing for a local search of new interaction patterns. Among the appealing features of the prior defined in this paper is the explicit consideration of edge direction and the possibility of feedback loops between proteins.

The large sample sizes characterizing RPPA data provide a unique opportunity for principled probabilistic modeling. At the same time, we are fully aware that departures from the multivariate Gaussian framework comes at a technical and computational cost. Particularly, computations involving partition functions, like the one reported in § 3.1 characterize an entire area of research (Møller et al. 2006, Friel et al. 2009) and they are known to be potentially problematic in high dimensional settings. The strategy suggested in our paper, relies on the ability to simulate from a multivariate binary distribution in a fast and reliable fashion. Therefore, if the number of proteins or samples is very large, one may consider alleviating the computation burden via parallelization across samples and/or relying on alternative approximation strategies (Besag 1974, Green and Richardson 2002; Friel et al. 2009). While a full review of this problem would be perhaps too ambitious for this application, we defer the reader to Zhou and Schmidler (2009) for a comparative discussion of alternative computational strategies.

On a related subject, it is worth noting that even standard Gaussian Markov random field representations are not fully immune from computational difficulties as closed form expressions are usually only available upon assuming very stringent restrictions (decomposability) on the graph topologies admitted for inference (Giudici and Green 1999, Wong et al. 2003, Roverato 2002, Atay-Kays and Massam 2005; among others). In contrast, the general Hammersley - Clifford representation (Hammersley and Clifford 1971, Besag 1974) does not prescribe unrealistic restrictions on the graph of interest and provides a very flexible recipe for the definition of conditional dependence relationships.

Finally, in this article we model dependence between protein assuming that there are no significant alterations in patterns of covariation between different subsets of patients. While this is beyond the scope of this work, our formulation provides a straightforward basis for methodological extensions aimed at allowing for formal tests of differential pathway activation, both in a supervised and unsupervised fashion. These possible developments will be particularly useful in the analysis of proteomic studies of AML, since the disease is markedly heterogeneous with numerous underlying genetic aberrations and we still lack a strategy for the treatment of different subtypes of AML based on current knowledge about genetic markers.

References

Airoldi, E. M., F. Markowetz, D. M. Blei, and O. Troyanskaya (2007). Statistical discovery of signaling pathways from an ensemble of weakly informative data sources. In *NIPS Workshop on Statistical Models of Networks*.

Andriew, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Markov chain Monte Carlo computation. *Annals of Statistics 37*(697-725).

Atay-Kays, A. and H. Massam (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika 92*(674-659).

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*(3), 870–897.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *(with discussion) Journal of Royal Statistical Society: Series B 36*, 192–236.

Boeuf, H., K. Merienne, S. Jacquot, D. Duval, M. Zeniou, C. Hauss, B. Reinhardt, Y. Huss-Garcia, A. Dierich, D. A. Frank, A. Hanauer, and C. Kedinger (2001). The ribosomal S6 Kinases, cAMP-responsive element-binding, and STAT3 proteins are regulated by different leukemia inhibitory factor signaling pathways in mouse embryonic stem cells. *The Journal of Biological Chemistry 276*, 46204–46211.

Chen, W., E. Drakos, G. Ioannis, E. Schlette, J. Li, L. Vasiliki, E. Staikou, E. Patsouris, P. Panayotidis, J. Medeiros, and R. GZ (2010). mTOR signaling is activated by FLT3 kinase and promotes survival of FLT3-mutated acute myeloid leukemia cells. *Molecula Cancer 9*, 292.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics 21*(113-120).

Dean, N. and A. E. Raftery (2005). Normal uniform mixture differential gene expression detection for cdna microarrays. *Bioinformatics 6*, 173–179.

Dobra, A., C. Hans, B. Jones, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis 90*, 196–212.

Efron, B. and R. Tibshirani (2006). On testing the significance of sets of genes. *The Annals of Applied Statistics 1*, 101–129.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science 303*, 799–805.

Friel, N., A. N. Pettitt, R. Reeves, and E. Wit (2009). Bayesian inference in hidden Markov random fileds for binary data defined on large lattices. *Journal of Computational and Graphical Statistics 18*(2), 243–261.

Giudici, P. and P. J. Green (1999). Decomposable graphical Gaussian model determination. *Biometrika 86 (4)*, 785–801.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82 (4)*, 711–732.

Green, P. J. and S. Richardson (2002). Hidden Markov models and disease mapping. *Journal of The American Statistical Association 97*(460), 1055–1070.

Hammersley, J. and P. Clifford (1971). Markov filed on fininte graphs and lattices. unplublished.

Hao, J., T. G. Li, X. Qi, D. F. Zhao, and G. Q. Zhao (2006). WNT/beta-catenin pathway up-regulates Stat3 and converges on LIF to prevent differentiation of mouse embryonic stem cells. *Dev Biol. 290*(1), 81–91.

Jain, N., T. Zhang, S. L. Fong, C. P. Lim, and X. Caoa (1998). Repression of Stat3 activity by activation of mitogen-activated protein kinase (MAPK). *Oncogene 17*(24), 3157–3167.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, J. Snyder, M. Greenblatt, and M. Gerstein (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 449–453.

Jones, B., C. Carvalho, A. Dobra, C. Hans, and M. West (2005). Experiments in stochastic computation for high–dimensional graphical models. *Statistical Science 20*, 388–400.

Kornblau, S., M. Minden, D. Rosen, S. Putta, A. Cohen, T. Covey, D. Spellmeyer, W. Fantl, U. Gayko, and A. Cesano (2010). Dynamic single-cell network profiles in acute myelogenous leukemia are associated with patient response to standard induction therapy. *Clinincal Cancer Research 16*(14), 3721–3733.

Kornblau, S., H. Vu, P. Ruvolo, Z. Estrov, S. O'Brien, J. Cortes, H. Kantarjian, M. Andreeff, and W. May (2000). BAX and PKCalpha modulate the prognostic impact of BCL2 expression in acute myelogenous leukemia. *Clinical Cancer Research 6*, 1401–1409.

Kornblau, S., M. Womble, Y. Qiu, E. Jackson, W. Chen, M. Konopleva, E. Estey, and M. Andreeff (2006). Simultaneous activation of multiple signal transduction pathways confers poor prognosis in acute myelogenous leukemia. *Blood 108*, 2358–2365.

Kornblau, S., H. Xu, W. Zhang, S. Hu, M. Beran, T. Smith, J. Hester, E. Estey, W. Benedict, and A. Deisseroth (1994). Levels of retinoblastoma protein expression in newly diagnosed acute myelogenous leukemia. *Blood 84*, 256–261.

Koster, J. T. A. (1996). Markov properties of non–recursive causal models. *Annals of Statistics 24*, 2148–2177.

Kreuter, M., K. Woelke, R. Bieker, C. Schliemann, M. Steins, T. Buechner, W. Berfel, and R. Mesters (2006). Correlation of neuropilin-1 overexpression to survival in acute myeloid leukemia. *Leukemia 20*, 1950–1954.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Clarendon.

Liang, F. (2010). A doubel metropolis-hastings sampler for spatial models with intractabel normalizing constants. *Journal of Statistical Computing and Simulation 80*, 1007–1022.

Martelli, A. M., C. Evangelisti, F. Chiarini, C. Grimaldi, L. Manzoli, and J. A. McCubrey (2009). Targeting the pi3k/akt/mtor signaling network in acute myelogenous leukemia. signaling network in acute myelogenous leukemia. *Expert Opin Investig Drugs 18*(9), 1333–1349.

Møller, J., A. N. Pettittt, R. Reeves, and K. K. Berthelsen (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalizing constants. *Biometrika 93*(2), 451–458.

Mrozek, K., G. Marcucci, P. Paschka, S. Whitman, and C. Bloomfiled (2007). Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? *Blood 109*, 431–448.

Mukherjee, S. and T. P. Speed (2008). Netrwork inference using informative priors. *Proceedings of the National Academy of Sciences 105*(38), 1433–14318.

Newton, M. A., F. Quintana, J. den Boon, S. Sengupta, and P. Ahlquist (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics in press.*

Palomero, T., M. Dominguez, and A. Ferrando (2008). The role of the PTEN/AKT pathway in NOTCH1-induced leukemia. *Cell Cycle 7*(8), 965–970.

Parmigiani, G., E. S. Garrett, R. Anbazhagan, and E. Gabrielson (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society B 64*, 717–736.

Pittman, J., E. Huang, H. Dressman, C.-F. Horng, C. S.H., M. Tsou, C.-M. Chen, A. Bild, E. Iversen, M. Liao, A. Huang, J. Nevins, and M. West (2004). Integrated Modelling of Clinical and Gene Expression Information for Personalized Prediction of Disease Outcomes. *Proceedings of the National Academy of Sciences 101*, 8431–8436.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms 9*, 223–252.

Roverato, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics 29*, 391–411.

Schäfer, J. and K. Strimmer (2005). A shrinkage approach to large scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology 4 (1)*(32).

Scott, J. and J. Berger (2008). Multiple testing, empirical Bayes, and the variable selection problem. *Duke University Department of Statistical Science, Technical Report* (2008–10).

Scott, J. and C. M. Carvalho (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics 17*, 790–808.

Sebastiani, P. and M. Ramoni (2005). Normative selection of Bayesian networks. *Journal of Multivariate Analysis 93*, 340–357.

Shojaie, A. and G. Michailidis (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology 16*(3), 407–426.

Shojaie, A. and G. Michailidis (2010). Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology 9*(1).

Tanner, S. M., J. L. Austin, G. Leone, L. J. Rush, C. Plass, K. Heinonen, K. Mrozek, H. Sill, S. Knuutila, J. E. Kolitz, K. J. Archer, M. A. Caligiuri, C. D. Bloomfield, and A. de la Chapelle (2001). BAALC, the human member of a novel mammalian neuroectoderm gene lineage, is implicated in hematopoiesis and acute leukemia. *Proceedings of the National Academy of Sciences 98*, 13901–13906.

Telesca, D., L. Inoue, M. Neira, R. Etzioni, M. Gleave, and C. Nelson (2009). Differential expression and network inferences through functional data modeling. *Biometrics 65(3)*, 793–804.

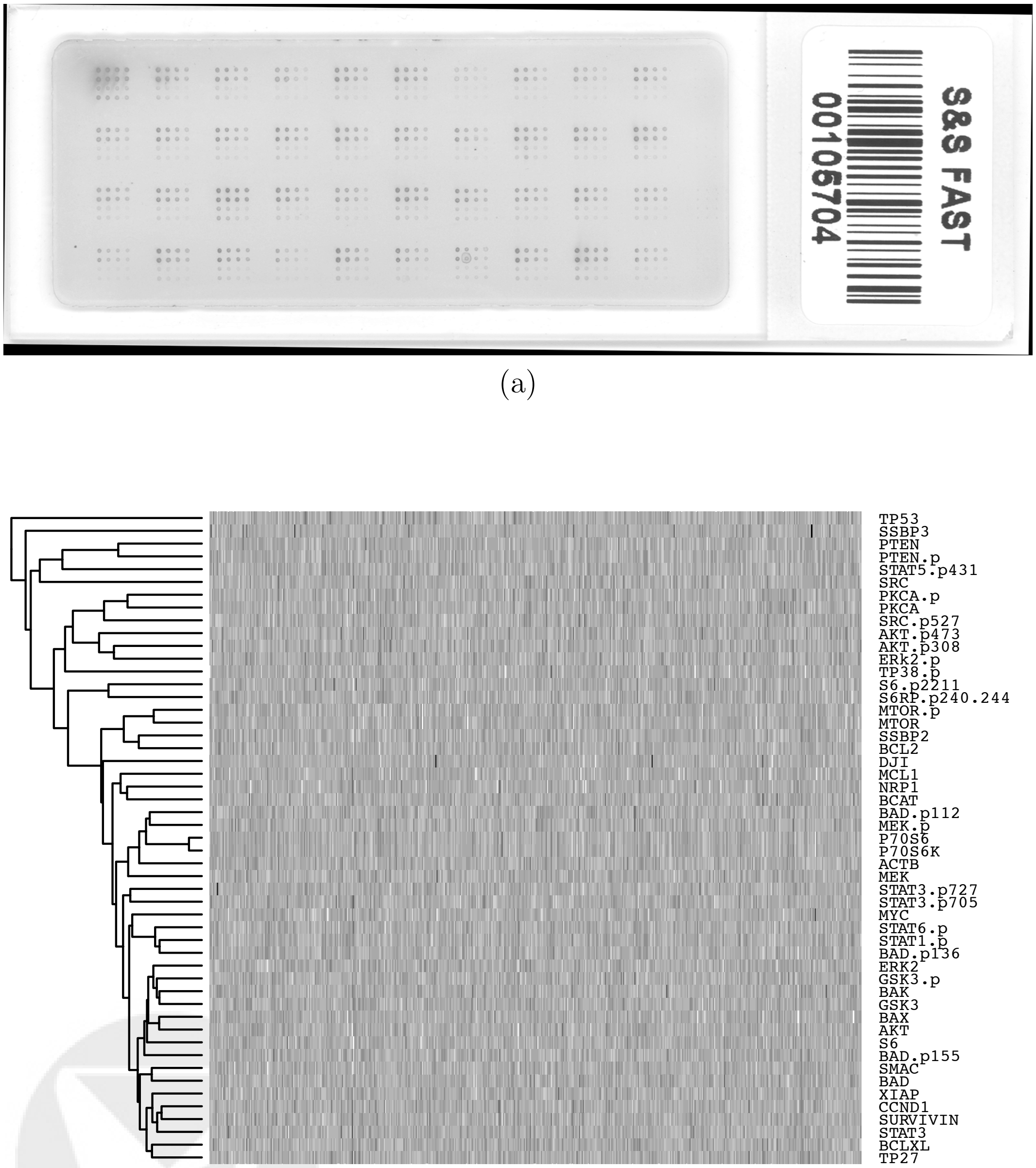
Telesca, D., P. Müller, G. Parmigiani, and R. S. Friedman (2010). Modeling dependent gene expression. Technical report, UCLA School of Public Health, Department of Biostatistics.

Tibes, R., Y. Qiu, B. Hennessy, M. Andreeff, G. Mills, and S. Kornblau (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cell. *Molecular Cancer Therapeutics 5*, 2512–2521.

Wan, X. and L. J. Helman (2003). Levels of PTEN protein modulate Akt phosphorylation on serine 473, but not on threonine 308, in IGF-II-overexpressing rhabdomysarcomas cells. *Oncogene 22*, 8205–8211.

Wong, F., C. K. Carter, and R. Khon (2003). Efficient estimation of covariance selection models. *Biometrika 90*, 809–830.

Zhou, X. and S. C. Schmidler (2009). Bayesian parameter estimation in Ising and Potts models: A comparative study with applications to protein modeling. Technical report, Duke University.



(a)

(b)

Figure 1: (a) A typical reverse phase protein array with 40 samples shown as the 40 batches on the slide. Each batch represents one individual sample with 16 spots, which are the results of duplicates of 8-step dilutions. (b) Normalized RPPA intensities for 51 proteins and 531 AML patients.

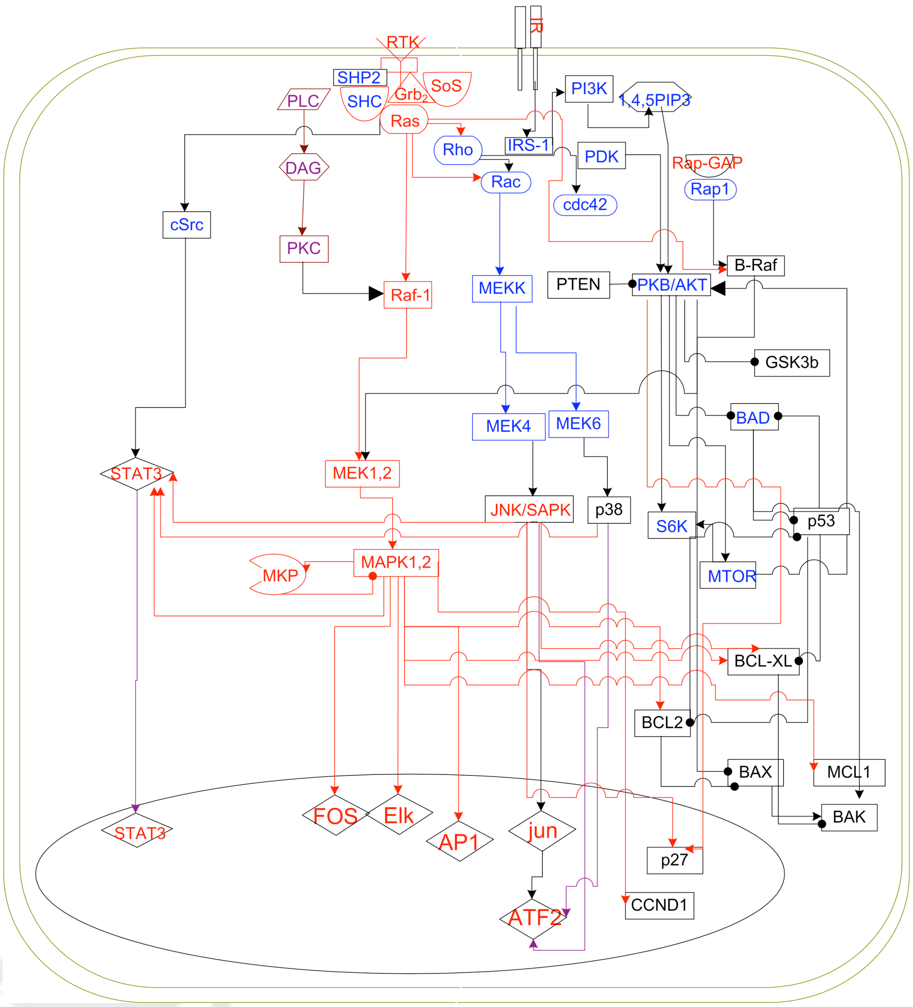


Figure 2: A protein interaction pathway produced by combining known protein-protein interactions from the literature. This network wiring diagram shows the connectivity of the receptor tyrosine kinase to the MAPK, AKT, STAT, BCL2,and p53 signaling proteins. We are able to measure a large percentage of the molecules using the RPPA for AML patients. The relationships between proteins suggested in this diagram will be considered as prior information for the proposed probability model.

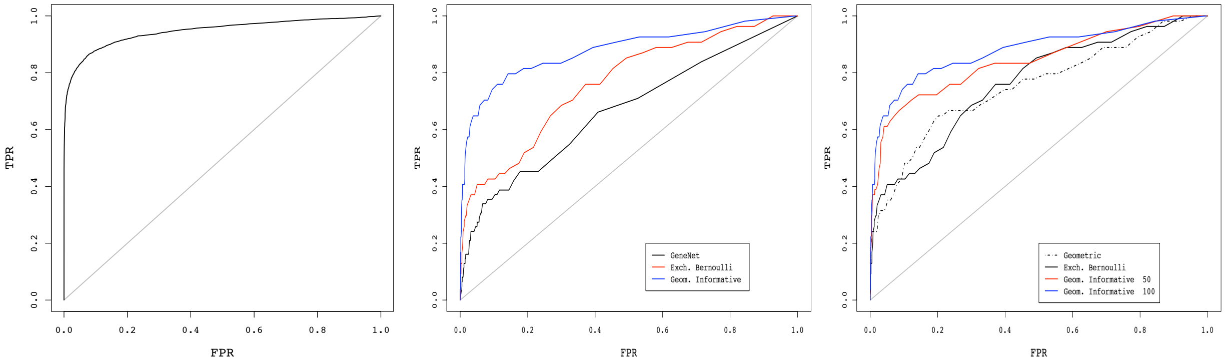


Figure 3: Simulated data. Panel (a) shows ROC curves for the correct classification of into activated versus non-activated proteins. Panels (b) and (c) show ROC curves for reporting edges as present or not. In panel (b) we compare different models. In panel (c) we compare alternative priors under the proposed model.

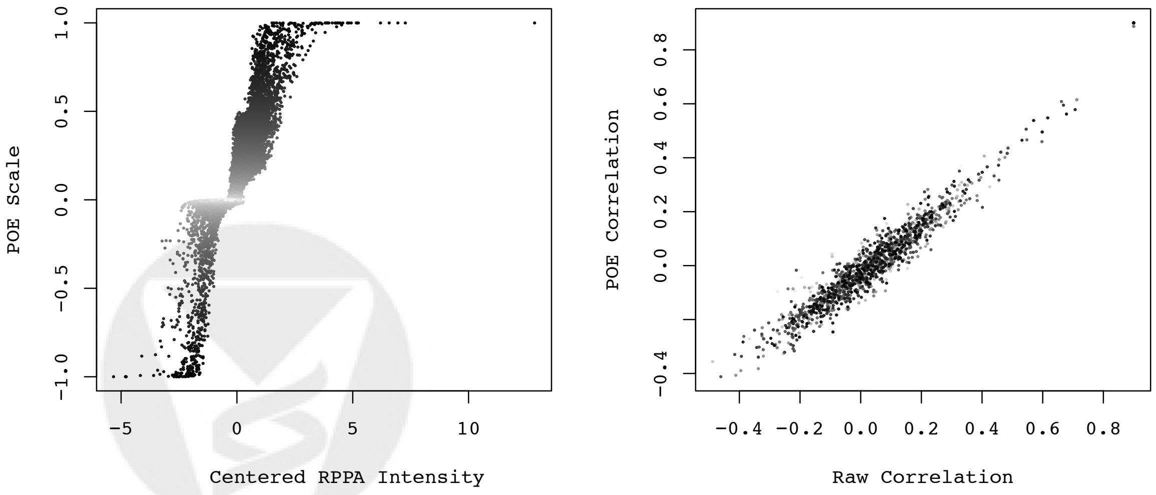
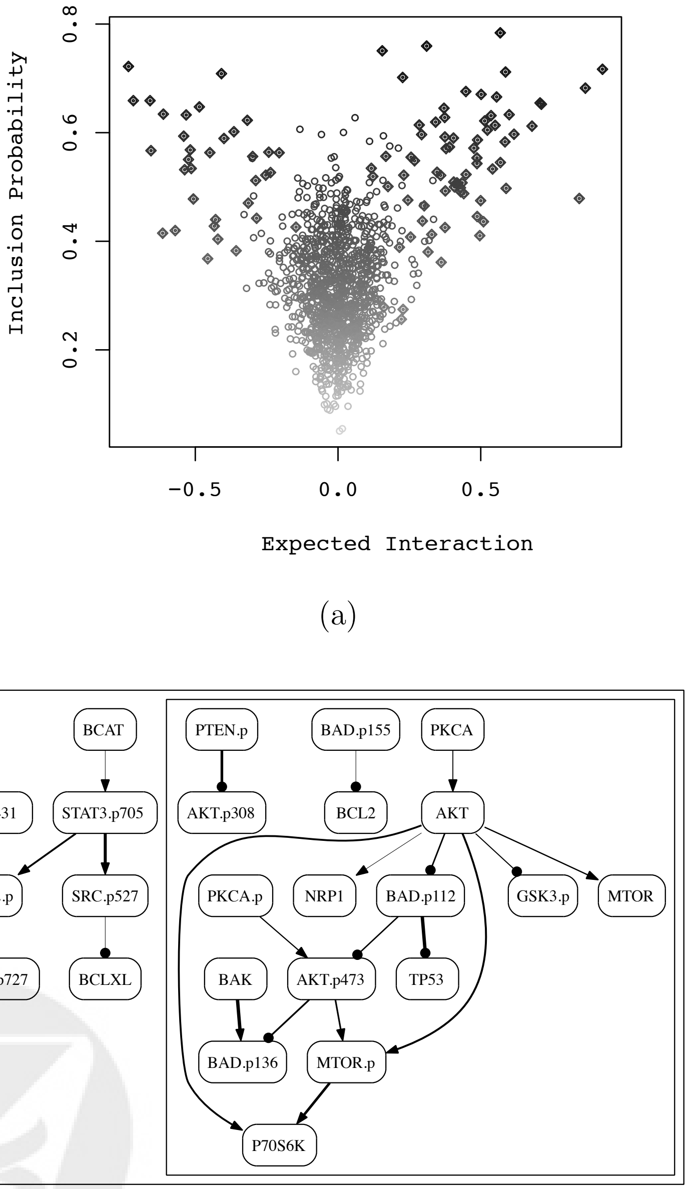


Figure 4: RPPA Study: *(Left Panel)* Centered protein abundance y_{gt} Vs. POE scale intensities $(p^*_{gt} = E(p^+_{gt} - p^-_{gt}))$. *(Right Panel)* Raw simple correlation estimates Vs. simple correlations in the POE scale.



(a)

(b)

Figure 5: *Panel (a):* Posterior expected interactions $E(\beta_u \mid \mathbf{Y})$ Vs. posterior edge inclusion probabilities $p((i,j)\ or\ (j,i) \in E \mid \mathbf{Y})$. Solid diamonds correspond to edges originally included in the prior pathway. *Panel (b):* Median model identified selecting edges with posterior inclusion probability greater than 0.5. Arrows define stimulatory relationships, whereas dotted arrowheads define inhibitory relationships. Edge thickness is proportional to the absolute size of the posterior expected interaction parameters.