Modeling Protein Expression and Protein Signaling Pathways

Donatello Telesca, Peter Müller, Steven M. Kornblau, Marc A. Suchard, and Yuan Ji

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 subpathways in relation to the unfolding of the biological process under study.

KEY WORDS: 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 article.

Statistical research in the estimation of dependence patterns has mainly centered on Bayesian networks (Jansen et al. 2003; Friedman 2004; 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 Michailidis 2009, 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, such as Gaussian Markov random fields (MRFs) (Wong, Carter, and Khon 2003; Jones et al. 2005; Scott and Carvalho 2008), 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

---

Donatello Telesca is Assistant Professor, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA (E-mail: dtelesca@ucla.edu). Marc A. Suchard is Professor, Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA, and Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA (E-mail: msuchard@ucla.edu). Peter Müller is Professor, Department of Mathematics, University of Texas at Austin (E-mail: pmueller@math.utexas.edu). Steven M. Kornblau is Professor, Department of Stem Cell Transplantation, University of Texas M. D. Anderson Cancer Center, Houston, TX (E-mail: skornbla@mdanderson.org). Yuan Ji is Director of Cancer Research Informatics, North Shore University Health System (E-mail: kauerzer@gmail.com).
proteins and activation states thought to occur in the human proteome. Furthermore, in typical protocols, one is not dealing with the $n \ll p$ 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; and (2) RPPA experiments often target specific biological 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 issue (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, suggested a similar approach in the context of Gaussian directed acyclic graphical (DAG) models. 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 causal perspective, we extend the work of Telesca et al. (2010) toward a multivariate probit link.

Conditional on \( e \), \( e_{gt} \). The observed abundance measurements \( y_{gt} \) 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 \( e_{gt} = (e_{g1}, \ldots, e_{gt}) \) is trinomial, with \( T \) possible trials, support \( \{-1, 0, 1\} \), and probabilities \( \pi_0 = p(e_{gt} = -1), \pi_1 = p(e_{gt} = 1) \), and \( \pi_2 = 1 - (\pi_0 + \pi_1) \).

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

We propose the following conditionally conjugate hierarchical normal prior for the remaining parameters in (2):
\[
\mu_g \sim N(m_\mu, \tau_\mu^2), \quad 1/\sigma^2 \sim \text{Ga}(\gamma_\sigma, \lambda_\sigma), \quad \alpha_i \sim N(0, \tau_\alpha)^2, \quad 1/k^- \sim \text{Ga}(\gamma_k, \lambda_k), \quad \text{and finally,} \quad 1/k^+ \sim \text{Ga}(\gamma_k, \lambda_k).
\]
Additionally, we will require \( \sum_{i=1}^7 \alpha_i = 0 \), for likelihood identifiability.

2.2 Modeling Dependence: \( p(e \mid \beta, G) \)

We model dependence between proteins using the formalism of graphical models (Lauritzen 1996). A graph \( G = (V, E) \) 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 the global Markov property associated with \( G \) (Besag 1974). Details, including whether edge set \( E \) is allowed to include directed edges, that is, ordered pairs of nodes, or only undirected edges, that is, 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., \( a \sim b \)), loops (e.g., \( a \rightarrow b, b \rightarrow c, c \rightarrow a \)), and reciprocal relationships (e.g., \( a \sim b, b \sim a \)), where \( a, b, c \in V \). Provided there are no directed
edges between vertices of the same undirected path (e.g., $a \rightarrow b$, $b \rightarrow c$, $c \rightarrow a$), this class of models has a clear causal interpretation and equivalent Markov classes may be defined via graphical moralization (Koster 1996; Lauritzen 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 $\mathcal{G}^m$. We proceed in the following way. We start with a prior on $\mathcal{G}$; this induces a prior on $\mathcal{G}^m$. Conditional on $\mathcal{G}^m$, we define a sampling model for the observed data. Finally, posterior updating provides the desired posterior inference on $\mathcal{G}$. In this process, it is important to note that we only learn about the features of $\mathcal{G}$ that would change $\mathcal{G}^m$. This process parallels inference in an analysis of variance (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 \(G^m\), standard 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 \(e_t = (e_{t1}, \ldots, e_{tG})\), \((t = 1, \ldots, T)\), which is consistent with the Markov structure spanned by a given reciprocal graph \(G\), through its moral representation \(G^m\). 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 \(z_t = (z_{g1}, g = 1, \ldots, G)\). The reduction to the multivariate binary vector \(z_t\) greatly simplifies this modeling step. Following the arguments by Hammersley and Clifford (1971) and Besag (1974), given \(G\) and the neighborhood structure encoded in a moral graph \(G^m\), the joint distribution \(p(z_t)\) can be defined in its full generality through the conditional joint probabilities \(p(z_{gt} | z_{(\neq t)})\). More precisely, let \(\mathbb{Q}(z_t) \propto p(z_t)\), then there exists an expansion of \(\mathbb{Q}(z_t)\), unique on \(\Omega = \{-1, 1\}^G\) and of the form:

\[
\mathbb{Q}(z_t) = \sum_{1 \leq g \leq G} z_{g1}H_g(z_{g1}) + \sum_{1 \leq g \leq G} \sum_{1 \leq h \leq G} z_{gh}z_{h1}H_{gh}(z_{gh}, z_{h1}) + \cdots + z_{1t}z_{2t} \cdots z_{Gt}H_{12\cdots G}(z_{1t}, z_{2t}, \ldots, z_{Gt}).
\]

The Hammersley–Clifford theorem (Besag 1974) postulates that while the \(H\)-functions may be chosen arbitrarily, for any set of labels \((1 \leq g, h, \ldots, s \leq G)\), \(H_{g,h,\ldots,s}\) may be nonnull if and only if the set of proteins labeled \((g, h, \ldots, s)\) forms a clique in the moral graph \(G^m\). 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 nonnull \(H\)-function with a single arbitrary parameter (Cox 1972; Besag 1974) and we can therefore write

\[
\mathbb{Q}(z_t) = \sum \alpha_s z_{s1} + \sum \beta_{gh} z_{gh} z_{h1} + \cdots + \beta_{12\ldots G} z_{1t} z_{2t} \cdots z_{Gt}.
\]

The foregoing formulation is a representation result for any multivariate distribution \(p(z)\). Later, conditioning on a particular graphical model \(G\) will substantially reduce the number of terms appearing in (4). In fact, when considering \(p(z_t | G)\), the function \(\mathbb{Q}(z_t | G)\) will only include sets of vertices forming cliques in the moralized graph \(G^m\). Finally, the definition of a joint distribution over the original trinary indicators \(e_t\) is completed with the conditional probability \(p(e_{gt} = -1 | z_{gt} = -1) = \exp(\gamma_{g1})/\{1 + \exp(\gamma_{g1})\}, (g = 1, \ldots, G)\).

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

2.3 Priors Over Interaction Parameters and Graphical Structures: \(p(\beta | G) p(G)\)

We introduce an informative class of priors for the unknown graph \(G\). 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 \(G_0 = \{V, E_0\}\). 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 \(p(G) \propto f(d(G_0, G))\), where \(d(\cdot, \cdot)\) is a discrepancy measure and \(f' \leq 0\), so that graphs that deviate from \(G_0\) are assigned a lower prior probability. A similar approach was introduced by Mukherjee and Speed (2008) for Gaussian DAGs.

Let \(A'\) denote the complement of set \(A\) and define \(d(G_0, G) = |E^c \cap E_0| + \delta [E \cap E_0^c]\), \(\delta > 1\). Here, \(|E|\) identifies the number of edges in a graph \(G = \{V, E\}\). This defines a discrepancy measure \(d(G_0, G)\) as the weighted sum of edges dropped from \(G_0\) and edges added to \(G\). If we assume for a moment \(\delta = 1\), then \(d(G_0, G)\) reduces to the number of changed edges relative to \(G_0\). A weight \(\delta > 1\) allows us to include a notion of parsimony by using an increased penalty for adding edges compared with dropping edges. For a given \(d(\cdot, \cdot)\), we assume that \(p(G | \varphi) \propto f(d(G_0, G)) = \varphi^{d(G_0, G)}\), (5)

where we assume \(\varphi \sim B(\alpha_\varphi, \beta_\varphi)\). The prior (5), with \(\delta = 1\), is equivalent to assuming exchangeable coin flips to determine each edge, with odds of matching edges in \(G_0\) equal to \(\varphi\). Similar probability models have been shown to provide in several contexts automatic multiplicity correction in the posterior (Scott and Berger 2008). For any given \(\varphi\), the expected density of a graph \(G\) is computed as \(\tilde{d}(G | \varphi) = (|E_0^c| - 1)/e_0^{\varphi} + |E_0|/e_0^{\varphi}\), which is a decreasing function of \(\delta\). Furthermore, the distribution of the graph density \(d(G | \varphi)\) has mode at \(\tilde{d}(G) = \tilde{d}(G_0)\) and variance decreasing with \(\delta\). A key property of (5) is therefore the informative nature of the prior, maximized at the prior guess \(G_0\), leading to posterior simulation being a local exploration of the model space. This is in sharp contrast to alternate 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 \(d(G_0, G)\). For example, this function could be defined in ways that force the inclusion of a subset of edges, say \(E_0^\ast\), or in ways that introduce discrete cutoffs to the overall model size.

Our model is completed by defining normal priors over the interaction parameters \((\beta_{gh}, \beta_{ghk}, \ldots)\) \((\forall g < h < k \ldots)\) so that \(\beta_{gh} | G \sim N(\eta_\beta, \sigma_\beta^2)\), for any generic index \(u\) of the parameter vector \(\beta\). The mean value \(\eta_\beta\) 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 labeled with an arrowed-tip \((i \rightarrow j)\) if interactions

\footnote{The density of a reciprocal graph \(G\) is defined as \(d(G) = \frac{|E|}{(|V|-1)!}\).}
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 Section 2.2 and focusing on second-order interactions parameters \( p_{gh} \). 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 \( p(z_{gt} = 1 | z_{ht} = 1) > p(z_{gt} = 1 | z_{ht} = -1) \). Similarly, for inhibitory relationships, we assume \( p(z_{gt} = -1 | z_{ht} = 1) > p(z_{gt} = -1 | z_{ht} = -1) \). It is easy to verify that these definitions motivate the choice of \( \eta_n > 0 \) for stimulatory and \( \eta_n < 0 \) for prior inhibitory interactions. In the absence of prior information, \( \eta_n \) is simply set to 0.

For reference, we summarize the complete model. Let \( \lambda_g = (k_g^-, k_g^+, s_g) \); we have:

Sampling model: 
\[
\gamma_{gt} | e_{gt} = e, \lambda_g \sim f_e(\gamma_{gt}),
\]
\[
\gamma_{gt} = y_{gt} - (\alpha_t + \mu_g)
\]

Autologistic:
\[
p(z_{1t} | G, \beta) \propto \exp \left( \sum_g \alpha_g z_{gt} + \sum_{g,h,t} E_{gh} p_{gh} z_{gt} z_{ht} \right)
\]

Graphical model prior: 
\[
p(G | \phi) \alpha \phi^Q(G, G)
\]

Hyperpriors:
\[
\beta_{ij} | G \sim N(\pi_{ij}, \sigma_{ij}^2), (i, j) \in E^m
\]
and \( \phi \sim \text{Beta}(\alpha, \beta) \),

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

In our analysis, we consider a default prior setting inspired by the considerations in Jiang (2007). Specifically, we choose \( \sigma_i^2 = 1/\epsilon_i^m, \phi \sim \text{Beta}(T, m, \xi), \) and \( \delta = \nu / \sqrt{m_G} \), where \( m_G = G(G - 1) \), and \( \xi > 0 \) is a tuning constant. This simplifies the posterior analyses to the choice of two parameters: \( \xi > 0 \), regulating overall sparsity, and \( \nu > 0 \), defining prior deviation penalties from \( \rho_{ij} \).

3. POSTERIOR COMPUTATION AND INFERENCE

3.1 Stochastic Search and MCMC Computation

We are interested in identifying patterns of molecular interactions, as informed by experimental data and guided by a prior pathway. The full model is determined by mixture membership indicators \( e \), a conditional dependence configuration summarized in a graph \( G \), MRF parameters \( \beta \), and a collection of parameters \( \theta \), which we assume contains all remaining random quantities with the exception of \( Y \). Our inference centers on the posterior distribution \( p(e, G, \beta, \theta | Y) \), 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 \( \theta, e, \beta, \text{and} G \). Updating of \( e \) and \( \theta \) can be implemented via standard Gibbs sampling.

Unfortunately, updating \( \beta \) and \( G \) remains problematic (Green and Richardson 2002). The complete conditional posterior for \( \beta \) and \( G \) depends on \( e \) only indirectly through \( z \). Recall that we further summarized the trinary indicators \( e \) into binary indicators \( z \in \{-1, 1\}^G \) of protein activation. The conditional posterior of \( \beta \) is given by
\[
p(\beta | Y, z, G, \theta) = p(\beta | z, G, \theta) \propto p(z | \beta, G, \theta) \pi(\beta | G, \theta).
\]
and the sum \( \sum_z \) is over all \( 2^G \) possible realizations of \( z \).

The full conditional distribution of \( \beta \) is therefore defined in terms of a partition function \( g(\beta) \), which requires summation over all possible \( 2^G \) realizations of \( z_t \), for any \( t = 1, \ldots, T \). 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 \( \beta \). Similar considerations apply for updates involving \( G \), with the further complication that changes in \( G \) define changes in the dimensionality of the parameter vector \( \beta \).

This problem is well known in the literature on MRFs and several solutions have been proposed, based on ad-hoc approximations 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 \( e \) and dependence parameters \( \beta \) in a joint fashion is possible without calculation of \( g(\beta) \).

3.2 Updating \( \beta \)

To update \( \beta \), we consider a Metropolis–Hastings (MH) transition kernel that involves changes in \( z \) and consequently in \( e \). For any generic element of \( \beta \), say \( \beta_u (u \in \{1, \ldots, M\}) \), we consider a proposal distribution \( (\beta_{u'}^*, e^* | \beta_u, e, G, \theta) \), where an arbitrary proposal density and possibly dependent on the previous values \( \beta_u \), whereas \( p(e | \beta, G, \theta) \) is the conditional prior defined in Section 2.2. In our implementation, for all \( i > j = 1, \ldots, G \), we consider a random-walk proposal \( q(\beta_{ij}) = \mathcal{N}(\beta_{ij}, s_p) \), where \( s_p \) is calibrated at burn-in to achieve acceptance rates between 30% and 70%.

Draws from \( e^* | \beta^*, G, \theta \) may then be obtained easily via a short Gibbs run, based on full conditionals \( p(e_{gt} | e, \beta, G, \theta) \). It is easy to verify that for the construction of a standard MH transition kernel the newly proposed values \( (e^*, \beta_{ij}^*) \) must be accepted with probability
\[
\rho(\beta_{ij}, e, e^*, \beta_{ij}^*) = \min \left\{ \frac{p(Y | e, \theta) \pi(\beta_{ij}^* | G, \theta) q(\beta_{ij})}{p(Y | e^*, \theta) \pi(\beta_{ij} | G, \theta) q(\beta_{ij}^*)}; 1 \right\}.
\]

The structure of the proposal distribution allows for the elimination of the partition function \( g(\beta) \), at the cost of having to simulate \( G \times T \) mixture indicators for each \( \beta_{ij} \). Gains in efficiency could be further achieved by considering block updates for the elements of \( \beta \). 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
We conclude by summarizing the overall MCMC. We use notation $x_i$ to indicate the vector $x$ without the element $x_i$, and $|y \mid z$ 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:

\[
[\theta \mid e, \beta, G^m, Y], \quad [e_{st} \mid e_{-st}, \beta, G^m, \theta, Y], \quad [\beta_{ij} \mid \beta_{-ij}, e, \theta, G^m]. \quad [G \mid e, \beta, \theta].
\]

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 alternate approaches. In brief, we simulated data under a model that differs from the proposed sampling model. Hypothetical data were sampled as: $Y_{ij} \sim \pi_{ij} N(-c, v^2) + (1 - \pi_{ij} - \pi_{ij}^2) N(0, 1) + \pi_{ij} N(c, v^2)$, where the mixing proportions depend on the realizations of a matrix normal random quantity $W \sim M(N(M, \Sigma, I_p)$ through a multivariate probit link (Chib and Greenberg 1998) so that $\pi_{ij} = p(w_{ij} < -1)$ and $\pi_{ij} = p(w_{ij} > 1)$. The concentration matrix $\Sigma^{-1}$ was structured via structural equations to encode a notion of true dependence in a reciprocal graph. We fitted our model using a misspecified network prior. Given the true set of edges $E^*$ (in our example, $|E^*| = 50$), we consider $E_0 = (E^* \setminus E_1^*) \cup E_1^*$, with $E_1^* \subseteq E^*$ omitted from the prior edge set (in our example, $|E_1^*| = 20$) and $E_1^*$ serving as a random misspecification set including false edges (in our example, $|E_1^*| = 30$). Ideally, this prior would mimic what is often the case in practice, where a prior pathway may include edges that do not play a relevant role in the biological process under investigation ($E_1^*$) as well as previously unidentified but relevant interactions ($E_1^*$).

Figure 3(a) explores robustness of protein activation inference to data arising from misspecified sampling models, by comparing inference on protein activation with the simulation truth. Figure 3(b) shows receiver operating characteristic (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. Figure 3(c) shows similar ROC curves under alternate priors with the proposed model, comparing the proposed informative prior (5) and two relatively noninformative priors. Details and sensitivity analyses are reported in the supplementary materials to this article, available online. Overall, we found that a sampling model mimicking a potential data generative process is likely to provide some advantages over standard multivariate Gaussian calculations. Furthermore, we remark how using prior information as a source of structural constraints on the model space can provide a substantial advantage in the identification of meaningful interactions.

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 process 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 the inference approach are 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 involved 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 preprocessing of data from any high-throughput experiments in general, and RPPA data in particular, is critically important. We used SuperCurve as it is 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 Section 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 $\pi_{gt} = p(e_{gt} = 1 \mid Y)$ and $\pi_{gt} = p(e_{gt} = -1 \mid Y)$, we follow Parmigiani et al. (2002) and define a probability of
expression (POE) 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 POE on the new scale ($p^*$). Plotting the raw correlation coefficients versus 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 toward large absolute values of the correlation coefficients.

Our inference on molecular interaction is based on posterior edge inclusion probabilities, $P((i, j) \in E \mid (j, i) \in E \mid Y)$. 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 $G$. In Figure 5, we represent the posterior edge inclusion probabilities for all possible protein interactions, against expected posterior interactions $E(\beta_{ij} \mid Y)$. Solid diamonds denote edges originally included in the prior pathway $G_0$. Our differential penalization scheme favors edges in the target pathway, but still allows for the model to explore interactions that 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 $E(\beta_{ij} \mid Y)$. We distinguish between stimulatory and inhibitory relationships. Following the argument introduced in Section 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 figure, we represent stimulation with arrows and inhibition with dotted arrowheads.

Our analysis identifies two main subpathways [Akt/mTOR (protein kinase B/mammalian target of rapamycin) and STATs (signal transducers and activators of transcription)] as groups of proteins exhibiting significant similarities in their overabundance 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 cell formation), and their upregulation via $\beta$-catenin (BCAT) has been documented by several authors (Boeuf et al. 2001; Hao et al. 2006). 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 in AML was also recently noted by Kornblau et al. (2010).

Some of our results are surprising. For example, pPTEN (phosphatase and tensin homolog) is inactivated when phosphorylated and therefore cannot inhibit Akt phosphorylation. The negative relationship between pPTEN and pAkt-308 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, e.g., the discussion in Palomero, Dominguez, and Ferrando 2008]. Likewise, one would expect the pAKTs to have an edge with pGSK-3 (glycogen synthase kinase 3) 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 (P70S6 kinase), 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 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 a great need for new therapeutic strategies.

We assess the sensitivity of our inference with respect to prior settings by repeating our analysis for a range of values associated with prior hyperparameters. We summarize the posterior distribution of the proteomic pathway by analyzing the degree of each protein. As we consider reciprocal graphs, the degree of a vertex (protein) is defined as the number of edges incident to the vertex, with reciprocal relationships counted twice. A summary of our results is reported in Figure 6, where we represent the posterior mean degree associated with each protein and 95% credible intervals. Our analysis considered varying
Figure 6. Posterior mean degree and associated 95% CI by protein. Panels (a), (b), and (c) correspond to increasing sparsity penalties $\varsigma = 0.5, 1.0,$ and $2.0$, respectively. For each protein, we report the posterior mean vertex degree evaluated in relation to three settings of the pathway deviation penalty parameter $\nu$ [$\nu = 0.5$ (black), $\nu = 1.0$ (blue), $\nu = 2.0$ (green)]. The online version of this figure is in color.
levels of prior sparsity ($\zeta = 0.5, 1.0, \text{ and } 2.0$), as well as varying degrees of prior pathway faithfulness, by considering pathway departure penalties, $\nu = 0.5, 1.0, \text{ and } 2.0$. We found that our results are robust to varying degrees of prior pathway faithfulness and overall sparsity. The proteomic degree distribution is maintained in agreement with respect to the order and magnitude of the expected number of neighbors. We observed some added instability associated with increasing levels of the sparsity parameter $\zeta > 2$. However, posterior degree distributions are, for the large majority, significantly overlapping for most proteins over wide ranges of prior hyperparameters.

6. DISCUSSION

We have proposed a probabilistic framework for the analysis of RPPA data. We focus on assessing the probability of protein–protein interactions by considering patterns of similarity characterized by hidden states of activation (as compared with 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 article is the explicit consideration of edge direction and the possibility of feedback loops between proteins.

We observed strong robustness of posterior inference about the pathway, to the setting of prior hyperparameters. This is expected as RPPA data are often characterized by large sample sizes. We however maintain that careful tuning and diagnostic checks are often required and are useful tools to assess sensitivity to the prior.

The large sample sizes characterizing RPPA data provide a unique opportunity for principled probabilistic modeling. At the same time, we are fully aware that departure from the multivariate Gaussian framework comes at a technical and computational cost. Particularly, computations involving partition functions, like the one reported in Section 3.1, characterize an entire area of research (Møller et al. 2006; Friel et al. 2009) and are known to be potentially problematic in high-dimensional settings. The strategy suggested in our article 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 alternate 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 alternate computational strategies.

On a related subject, it is worth noting that even standard Gaussian MRF 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; Roverato 2002; Wong, Carter, and Khon 2003; 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 by 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 in an 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 the current knowledge about genetic markers.

SUPPLEMENTARY MATERIALS

The supplementary materials are intended to provide an expanded discussion of some of the aspects characterizing the methodology proposed in our article. We discuss identifiability of the sampling model’s mixture components and assess how the proposed methodology compares to current network inference techniques. We also provide full conditional distributions, when available.

[Received October 2010. Revised May 2012.]

REFERENCES


Downloaded by [University of California, Los Angeles (UCLA)] at 11:55 29 December 2012