Sparse Graphical Models for Exploring Gene Expression Data

Adrian Dobra[†◇‡] [*]
Chris Hans[◇‡], Beatrix Jones[◇‡], Joseph R Nevins[†] & Mike West[◇]

Abstract

We discuss the theoretical structure and constructive methodology for large-scale graphical models, motivated by their potential in evaluating and aiding the exploration of patterns of association in gene expression data. The theoretical discussion covers basic ideas and connections between Gaussian graphical models, dependency networks and specific classes of directed acyclic graphs we refer to as compositional networks. We describe a constructive approach to generating interesting graphical models for very high-dimensional distributions that builds on the relationships between these various stylized graphical representations. Issues of consistency of models and priors across dimension are key. The resulting methods are of value in evaluating patterns of association in large-scale gene expression data with a view to generating biological insights about genes related to a known molecular pathway or set of specified genes. Some initial examples relate to the estrogen receptor pathway in breast cancer, and the Rb-E2F cell proliferation control pathway.

Key words: Bayesian regression analysis; Compositional networks; Estrogen receptor gene and pathway; ER pathway; Gene expression; Graphical models; Model selection; Rb-E2F genes and pathway; Transitive gene expression pathways

1 Introduction and Perspectives

The interest in exploratory methods for evaluating patterns of association between many variables has been invigorated by problems involving large-scale gene expression studies. Problems of modeling structure under the "large p, small n" paradigm [1] challenge modern statistical science both conceptually and computationally. We discuss some aspects of this here in the context of exploratory uses of large-scale Gaussian graphical models in evaluating patterns

[*] *Corresponding author* ISDS, Duke University, Durham NC 27708, USA. *email:* adobra@isds.duke.edu, *tel/fax:* +1 (919) 684-8795/8594
[†] Department of Molecular Genetics & Microbiology, Duke University
[◇] Institute of Statistics & Decision Sciences, Duke University
[‡] Statistical & Applied Mathematical Sciences Institute, RTP

of association with a view to generating biological insights plausibly related to underlying biological pathways. Our outlook is similar to that of [2] where direct, correlation based analysis was able to define biologically interpretable connections transitively across groups of genes. In contrast, however, we are concerned here with a consistent theoretical framework, and a constructive methodological approach, to generating proper graphical models - representing the complete joint distributions - in tens of thousands of dimensions.

Specifically, we develop methods for modeling the covariance structure of high-dimensional distributions with a focus on *sparse* structure – particularly relevant in modeling associations in gene expression data. We develop a constructive approach to generating large-scale undirected Gaussian graphical models based on representation of the joint distribution of variables via sets of linear regressions. Key foci include questions of consistent prior specifications using an encompassing Wishart prior for precision matrices (inverse covariance matrices) in undirected graphs. The relevance of formulations in terms of directed acyclic graphs and complete conditional distributions/dependency networks are also discussed and utilized.

The notion of sparsity of graphical models for gene expression data reflects the view that patterns of variation in expression for a gene, G, will be well-predicted by those of a relatively small subset of other genes. Beyond parsimony, this embodies a view that transcriptional regulation of a single gene or pathway component is generally defined by a small set of regulatory elements. Our theoretical framework reflects this, providing the first statistical approach to large-scale but *sparse* graphical models in which each variable/gene is expected to have a small number of direct neighbors.

The scope for application of the statistical analysis here is explicitly that of randomly sampled, complete observational data. Graphical models are simply models of joint distributions assuming a random sampling paradigm. We are *not*, here, concerned with developing models of causal gene networks; that requires a context of experimentation and intervention to understand directional influences, rather than our observational, random sampling paradigm. Directed graphical models (Bayesian networks – directed acyclic graphs or DAGs) have explicit causal modeling goals, but our use of DAGs here is purely technical; a DAG formulation provides a natural and useful technical step in the identification of high posterior probability undirected graphical models for randomly sampled data.

The specific algorithmic approach to generating high-posterior probability graphs described here utilizes forward/backward variable selection within sets of linear regressions that define the full joint graphical model. This is an initial approach that demonstrates the utility of the conceptual framework of the regression model formulation via DAGs and its feasibility in very high-dimensional problems; we are able to efficiently generate such models with thousands of variables, going far beyond what has been contemplated in the literature to date. We extend this to generate multiple candidate graphical models, and this underlies our examples.

Our applied perspective is that these models have utility as exploratory tools. Graphical representations of these, and other, models provide displays of gene expression based information that may be explored to generate insights about pathways. Our examples illustrate this. The fitting of complete multivariate distributions moves us beyond the usual correlation-based approaches, such as clustering approaches, that aim to identify groups of genes with related patterns of expression. In a graphical model, the first-order neighbors of a specific gene G are those genes that, conditional on the underlying model, together provide the set of predictors of gene expression variation in G. They render G conditionally independent of all other genes. Some genes identified as neighbors of G may be only very weakly correlated with G, but highly related in terms of partial correlation in the context of the other neighbors, i.e., in the regression sense defined implicitly by the graphical model. Such gene-gene relationships would simply not be "discovered" by correlation-based methods, and it is clear that they may define significant relationships in a biological pathway. Thus, in addition to the exploration of transitive relationships between pairs or sets of genes based on short paths in interesting graphs, a key applied interest in these models lies in this capacity to expand our ability to identify candidate genes that may play roles in a specific pathway under investigation. These points are highlighted in our example concerning the Rb-E2F cell growth and proliferation control pathway; this represents an example of initial applied investigations to begin the process of moving the theory of graphical models into applications in more than a small number of dimensions.

2 Graphs, Networks and Regressions

We are interested in models of the patterns of dependence in p-dimensional normal distribution. Write x for a random normal $p-$vector. Assume centered data measured on a common scale, and denote the model by $x \sim N_p(0, \Sigma)$ with positive definite variance matrix Σ and precision matrix $\Omega = \Sigma^{-1}$. Denote the elements of Ω by ω_{ij}, and those of Σ by σ_{ij}.

In our applications x is a vector of gene expression levels and p is very large; the example has $p = 12,558$. Write x_i for the i^{th} univariate element of x and, for any subset of indices $K \in \{1, \ldots, p\}$, write $x_K = \{x_i : i \in K\}$. Two key subsets are $[-i] = \{1, \ldots, p\} \setminus \{i\}$ and $k : m = \{k, k+1, \ldots, m\}$ for any $k \leq m$. The joint distribution implies the set of consistent complete conditional univariate distributions $p(x_i | x_{[-i]})$ for $i \in \{1, \ldots, p\}$. These are univariate normal linear regressions with partial regression coefficients of x_i on x_j given by ω_{ij}/ω_{ii}, $(j \in [-i])$.

2.1 Sparse Gaussian Graphical Models

Write $ne(i) = \{j : \omega_{ij} \neq 0\}$, so that the variables $x_{ne(i)}$ are the predictors of x_i – the variables that render x_i conditionally independent of all other x_k for $k \notin ne(i)$. Thus $p(x_i|x_{[-i]}) = p(x_i|x_{ne(i)})$. In graph theoretic and graphical modeling terms, $x_{ne(i)}$ defines the neighbors (or Markov blanket) of x_i [3]. Our focus is on sparse models, with each such conditional regression involving a relatively (to p) small set of predictors, so that each row (hence column) of Ω will have many zero entries; thus $\#ne(i)$ will be small. This is consistent with the view that, though a gene may play a role in many biological pathways and be associated with many other genes, the "influence" of many other genes will often be transitively represented by a small number of directly "predictive" genes in a small neighborhood of x_i, consistent with sparsity of Ω.

To connect more properly with (undirected) graphical models, write $V = \{1, \ldots, p\}$ and define the *graph* $\mathcal{G} = (V, E)$ over *nodes* V by an *edge set* E such that an edge exists between nodes i and j if, and only if, $j \in ne(i)$ (hence, by symmetry, if and only if $i \in ne(j)$). The normal distribution $p(x)$ defines a graphical model over the graph $\mathcal{G}$ [4] – a proper joint distribution for x whose conditional independence structure is explicitly exhibited through the edge sets characterizing the $p(x_i|x_{ne(i)})$.

2.2 Dependency Networks

The challenge of fitting high-dimensional distributions may be addressed by fitting sets of univariate conditional distributions. This motivated the general *dependency networks* of Heckerman *et al* [5]. A dependency network is simply a collection of conditional distributions $\{p(x_i|x_{[-i]})\}$ that are defined and built separately. In the specific context here, of sparse normal models, these would define a set of p separate conditional linear regressions in which x_i is regressed on a small selected subset of other variables, say $x_{pv(i)}$ for some **p**redictors **v**ariables $pv(i)$, with each being determined separately. Generally, of course, these distributions will not cohere in the sense of being consistent with a proper joint distribution, and this is almost surely the case in problems with many variables where variable selection and collinearity issues will lead to inconsistent models. Even the basic consistency requirement $j \in ne(i)$ if, and only if, $i \in ne(j)$ will generally be violated for many i, j.

Nevertheless, it is most attractive and useful to explore high-dimensional data relationships by fitting – via whatever means – sets of univariate regression models $\{p(x_i|x_{pv(i)})\}$ in which $pv(i)$ is selected as part of the univariate modeling process. This can then be used as an initial step in generating relevant and consistent models of the full joint distribution, as we show below.

2.3 Compositional Networks

A natural, direct route to specifying a set of univariate models that cohere is via a "triangular" model that defines the joint distribution by composition, i.e.,

$p(x) = \prod_{i=1}^{p-1} p(x_i|x_{(i+1):p})\ p(x_p)$. This requires a specific ordering of variables, which we take here for discussion as simply $1, 2, \ldots, p$ with no loss of generality.

Suppose we now fit the set of p univariate regressions implied here, selecting specific subsets of **c**ompositional **p**redictor **v**ariables $cpv(i) \subseteq \{(i+1) : p\}$ for each i (and with $cpv(p)$ empty), and choosing regression parameters and conditional variance values for each i. This defines a set of models

$$x_i = \sum_{j \in cpv(i)} \gamma_{ij} x_j + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim N(0, \psi_i)$. The composition of the joint distribution $p(x)$ can then be written in structural form as

$$x = \Gamma x + \epsilon, \qquad \epsilon \sim N_p(0, \Psi), \tag{2}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_p)'$, $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$, and Γ is the $p \times p$ upper triangular matrix with zero diagonal elements and upper triangular, non-diagonal entries $\Gamma_{ij} = \gamma_{ij},\ (j > i)$.

Connecting with graphical models, define the directed acyclic graph (DAG) over nodes V through edges E that are directed from each element of $cpv(i)$ to i for all $i \in V$. By construction this graph is acyclic, and the joint distribution $p(x)$ defined by composition represents a Bayesian network over the DAG [4]. The sequence of ordered nodes $i = 1, \ldots, p$ defines a *well-ordering* of the vertices underlying this network [6]. Further, the structure corresponds to an underlying undirected graph that is obtained directly by replacing all arrows by undirected edges, and "marrying" all pairs of parents of any variable by joining them with an edge. The resulting edges correspond to the non-zero elements of Ω – nodes i, j are joined by an edge if, and only if, $\Omega_{ij} \neq 0$. It is this undirected graph – that directly defines the set of neighbors of any node and exhibits the conditional independence structure of the implied joint distribution – that is of key interest in describing and exploring associations among variables.

Moving to inference on the regressions, write $\theta_i = (\gamma_i, \psi_i)$ where γ_i is column vector of regression coefficients $\{\gamma_{ij} : j \in cpv(i)\}$, and set $\Theta = \{\theta_i : i \in V\}$. Then, including parameters in the notation, we have $p(x|\Theta) = \prod_{i=1}^{p-1} p(x_i|x_{(i+1):p}, \theta_i)\ p(x_p|\theta_p)$ (and note that $\theta_p = \psi_p$.) Fitting the set of regression models may be performed independently, in parallel, assuming priors over regression parameters that are independent across equations, or using reference priors and likelihood-based approaches. Prior independence across equations (or global independence on the DAG [7]) implies posterior independence and hence separate, parallel analyses provide the full joint posterior for the θ_i. Aspects of prior specification, and the issues of variable selection to define $cpv(i)$ within each equation, are discussed below.

The reduced form of the Bayesian network (2) is immediate: $x = (I - \Gamma)^{-1}\epsilon$

so that $\Omega = (I-\Gamma)'\Psi^{-1}(I-\Gamma)$. For any set of parameter values Θ we can then compute Ω and, by inversion, Σ. This computation is substantially eased by noting that we have immediate access to a sparse Cholesky decomposition component of $\Omega = LL'$ where L is the sparse lower-triangular matrix $L = (I-\Gamma)'\Psi^{-1/2}$. This then provides direct access to variance-covariances of defined subsets of variables via the solution of two lower-triangular systems of equations.

3 Analysis and Selection of Regressions

Focus now on any univariate regression equation for a chosen x_i, with k_i predictors in a predictor variable index set $pv(i) \subseteq [-i]$. That is, $(x_i|x_{pv(i)}, \theta_i) \sim N(x'_{pv(i)}\gamma_i, \psi_i)$ where $\theta_i = (\gamma_i, \psi_i)$. For discussion in this section, we look at the general case when the predictor set may involve any of the $p-1$ variables $x_{[-i]}$. In later sections, these results then apply to specific choices of $pv(i)$ that arise in dependency networks (where the elements of $pv(i)$ can be any values in $[-i]$), and then compositional networks (where, given the defined ordering of variables, the elements of $pv(i)$ are restricted to values in $(i+1):p$).

Prior specification may be guided by contextual information, as available, and general statistical principles. One consideration is scale; in gene expression studies, the x_j are measured on a common scale (usually log expression). Also, for exploratory data analysis, we will generally use priors that treat variables exchangeably rather than aim to incorporate substantive prior information – a key goal being to generate insights and clues about underlying biological interactions from a relatively unbiased initial standpoint. This implies priors that treat the elements of γ_i exchangeably.

A key statistical principle is the consistency of models and priors as the number of predictor variables, k_i, changes. We are interested in comparing models with different candidate predictor sets $pv(i)$, and each model will have its own prior for the implied θ_i parameters. We define consistent priors on subsets by embedding in an overall encompassing model/prior, as follows.

3.1 Regression Parameter Priors

Return to the full set of p variables $x \sim N_p(0, \Sigma)$. Any specified index set $a(i) = (i, pv(i))$ defines the corresponding θ_i as a transformation of elements of Σ. Hence a prior for Σ, or equivalently its inverse Ω, defines consistent priors on any regression model parameters for any chosen indices. The selection of subsets can then be governed by a variable selection prior that defines prior inclusion/exclusion probabilities for variables. In the encompassing model, we utilize the Wishart/inverse Wishart prior structure that is simple and effective in terms of both the statistical features (scale, exchangeability) and in terms of computation. We use the notation for inverse Wishart models of Dawid [8] (see also [9] and, with somewhat different notation, [10]). Begin with an inverse Wishart prior [8,11] $\Sigma \sim IW_p(\delta, \tau I)$ with δ degrees of freedom and scale matrix τI_p, for some $\tau > 0$. For small δ (and in this notation, the constraint is only that $\delta > 0$), this is a diffuse prior that, through the diagonal scale matrix, shrinks towards lack of correlation between variables and respects the common scale of measurement.

Now, the variable subset $x_{a(i)} = (x_i, x_{pv(i)})$ has variance matrix $\Sigma_{a(i)}$ from the relevant components of Σ, and the prior implies that $\Sigma_{a(i)} \sim IW_{k_i+1}(\delta, \tau I_{k_i+1})$. Partition $\Sigma_{a(i)}$ as

$$\Sigma_{a(i)} = \begin{array}{c} 1 \\ k_i \end{array} \begin{bmatrix} \sigma_{ii} & \kappa_i' \\ \kappa_i & \Sigma_{pv(i)} \end{bmatrix}.$$

From standard normal theory, the regression is $x_i \sim N(x_{pv(i)}' \gamma_i, \psi_i)$ where $\gamma_i = \Sigma_{pv(i)}^{-1} \kappa_i$ and $\psi_i = \sigma_{ii} - \kappa_i' \Sigma_{pv(i)}^{-1} \kappa_i$. Using standard distribution theory [10,11] we deduce the implied prior of the normal/inverse gamma form:

$$(\gamma_i | \psi_i) \sim N_{k_i}(0, \tau^{-1} \psi_i I_{k_i}) \quad \text{and} \quad \psi_i^{-1} \sim Ga((\delta + k_i)/2, \tau/2). \tag{3}$$

These priors are consistent with the encompassing prior on the full covariance structure of all variables, so we are working in a consistent framework as we move across equations i and subsets $pv(i)$ within equations.

3.2 Posterior Distributions and Marginal Model Likelihoods

Consider now the posterior and predictive aspects of analysis on processing n observations. For notation, write y_i for the $n-$vector of observed values of x_i, and Z_i for the $n \times k_i$ design matrix defined by the corresponding n observations on the predictor vector $x_{pv(i)}$. The vector form is simply $y_i \sim N_n(Z_i \gamma_i, \psi_i I_n)$ and the posterior and predictive distributions follow in standard normal/inverse gamma form [12]:

- $(\gamma_i | \psi_i, y_i, Z_i) \sim N_{k_i}(M_i^{-1} Z_i' y_i, \psi_i M_i^{-1})$ where $M_i = \tau I_{k_i} + Z_i' Z_i$; and
- $(\psi_i^{-1} | y_i, Z_i) \sim Ga((\delta + k_i + n)/2, (\tau + q_i)/2)$ where $q_i = y_i' y_i - y_i' Z_i M_i^{-1} Z_i' y_i$.

To compare models based on different subsets of predictor variables we require the value of the marginal model likelihood, i.e., the observed value of the prior predictive density function $p(y_i | Z_i)$. This follows, after some manipulation (see also [12,9]) as

$$p(y_i | Z_i) \propto \frac{\tau^{k_i/2} \, \Gamma((n + \delta + k_i)/2)}{|M_i|^{1/2} \, \{1 + q_i/\tau\}^{(n+\delta+k_i)/2} \, \Gamma((\delta + k_i)/2)}. \tag{4}$$

We do need to consider the special cases of the null model – when there are no predictors. Formally, $k_i = 0$ and γ_i is null. The results are as above, but with $k_i = 0$ and $q_i = y_i' y_i$.

3.3 Variable Subsets and Inclusion Probabilities

Our interest in sparse models recommends priors on subsets of variables $pv(i)$ in the compositional network that are penalized for larger values of $k_i = \#pv(i)$. A traditional Bayesian variable selection prior, specified in terms of a common inclusion/exclusion probability, is a first step, and the basis of analysis in examples explored here. Assuming a common inclusion probability of β, the prior probability of a model on a specific set of k predictors relative to that on a different, though possibly overlapping, set of $k+h$ specified predictors is just $(1-\beta)^h\beta^{-h}$. To be effective in problems with many candidate predictors the variable-inclusion probability must be very small. The implied binomial prior on the number of predictors in the regression can be used to assess this. With r candidate predictors, for β values smaller than $1/r$ the prior mass on $k=0$ increases, to a value of, for example, about 0.6 if $\beta = 0.5/r$ and around 0.9 at $\beta = 0.1/r$. These values give crude guidelines as to what "small" means relative to the number of available predictors. Smaller β values more strongly penalize complexity and induce sparsity of the resulting compositional network.

With priors specified we can search the model space stochastically or deterministically. It is understood that generating search or sampling methods for even moderately large values of p is a major challenge and a currently active research area. Our study and example here are based on forward/backward variable selection to identify a set of locally optimal regressions, one for each x_i, that define local modes in regression model space for each i. That is, for each i, we apply forward/backward selection with respect to the defined set of candidate predictors, and score each regression by the posterior model probabilities; for a regression on k parameters from a candidate set of r, the unnormalized posterior probability is simply $\beta^k(1-\beta)^{r-1-k}$ multiplied by the model marginal likelihood of equation (4).

4 Constructing Graphical Models from Regressions

Our constructive approach to generating sparse graphical models builds on the simplicity of dependency networks and univariate regressions just described. In summary:

- We build sets of regressions in an initial dependency network framework, and use these to generate an appropriate ordering of the variables to underlie a compositional network;
- With this ordering, variable selection for each linear model for the ordered variables produces a DAG;
- This DAG implies a unique undirected graph, and posterior distributions for the sets of regression parameters in the DAG then provide the means for inferences in the undirected graph.

Key to this approach is an understanding of the evaluation of posterior graph probabilities arising from DAGs, and the intuition behind the specific method of ordering variables to define a DAG. Let $\{(x_i|x_{pv(i)}) : i = 1, 2, \ldots, p\}$ be a dependency network obtained by maximizing the posterior regression probabilities $p(x_i|x_{pv(i)}) \cdot (\beta/(1-\beta))^{\#pv(i)}$, $i = 1, 2, \ldots, p$. Define the score associated with this dependency network to be the product:

$$\prod_{i=1}^{p} p(x_i|x_{pv(i)})(\beta/(1-\beta))^{\#pv(i)}. \tag{5}$$

Were the dependency network to be, in fact, a compositional network, then the score in (5) would be proportional to the posterior probability of the implied DAG model. Geiger and Heckerman [10] prove that, under priors for the regression parameters induced by the inverse Wishart prior on the covariance matrix Σ, any two DAGs in the same equivalence class [13] have the same marginal likelihood. Moreover, two such equivalent DAGs have the same number d of directed edges and hence they have the same prior probability $(\beta/(1-\beta))^d$; hence they each have posterior probability proportional to (5).

In a dependency network, the predictors for each variable x_i can be selected from all the other variables. A compositional network restricts this choice: candidate predictors for x_i are restricted to the variables that succeed x_i in the ordering. As a result, the score in (5) represents an upper bound for the unnormalized posterior probability of the DAG resulting from a compositional network $\{(x_i|x_{cpv(i)}) : i = 1, 2, \ldots, p\}$. The closer the dependency network is to a DAG, the closer this score will be to the required posterior probability. This underlies the use of the score to define the ordering of variables in our DAG: the ordering aims to to maximize the resulting posterior probability on the resulting model in an iterative process that, at each step, decreases the overall score (5) the least. The construction is now fully described.

4.1 Model Construction

Dependency network initialization

For each i, find a regression model $(x_i|x_{pv(i)})$ by search over all possible variables $pv(i) \subseteq [-i]$. Use forward/backward selection method (for example) to find local posterior modes in regression model space with predictor sets $pv(i)$.

Set iterate counter $h = 0$. Initialize an *ordered variable index vector* O as the null vector, and the *candidate variable* index set $C = V = \{1, \ldots, p\}$. Assign each x_j, $j \in C$, the *explanatory score*:

$$s_j^O = \prod_{i \in C} p(x_i|x_{pv(i)\setminus\{j\}})(\beta/(1-\beta))^{\#pv(i)-1}.$$

The earlier discussion motivates the use of this score on theoretical grounds, and we note also that, the smaller the score, the greater is the "importance" of this variable in explaining the other variables. The score of x_j can be easily obtained from the score of the current dependency network $\{(x_i|x_{pv(i)}) : i \in C\}$

by re-calculating the posterior probability of the regressions in which x_j is an explanatory variable.

Iterative ordering of variables and compositional DAG generation

For $h = 1, 2, \ldots, p-1$, repeat the following.

(1) Choose the candidate variable with highest score, $j_h = argmax_{j \in C}\{s_j^O\}$.
(2) Variable x_{j_h} becomes the next in the well-ordering, recorded by an update of the ordered variable index vector from O to $[O, j_h]$. Record $cpv(j_h) = pv(j_h)$ for this variable.
(3) Remove j_h from the candidate variable index set by an update of C to $C \setminus \{j_h\}$.
(4) For all $i \in C$ such that $j_h \in pv(i)$, reselect and fit the regression model for x_i choosing potential predictors from C only, to generate possibly revised models. These models then define updates of the sets $pv(i)$ and explanatory scores s_i^O for all $i \in C$.

At $h = p$, there is a single variable remaining in C that completes the ordered list O. The result is a compositional network defined by O and the index sets $cpv(i)$, i.e., a Bayesian network over the resulting DAG. We can now extract inferences about various aspects of the model, including regression coefficient estimates and, by transformation, estimates of Ω and hence whatever components of Σ are deemed of interest. Finally, by removing arrows and marrying all parents of any node in the graph we deliver the resulting "moralised" graph – the undirected graph, with the associated unnormalized posterior probability.

4.2 Sampling to Generate Multiple Graphs

This specific construction, employing forward/backward selection methods for finding regression models as well as the iterative procedure for variable ordering, is deterministic. It is a so-called greedy deterministic approach and, as such, is naturally sensitive to data perturbations that would lead to other final models. Clearly the ideas here are general and the specific construction represents an initial approach that could be modified and extended in a number of ways. Generating multiple candidate regression models for each variable in the compositional network allows us to evaluate multiple resulting DAGs, hence undirected graphs, and compare and contrast them, and this can easily be embedded in the DAG generation step. It is critical to begin to explore multiple models to understand model uncertainty and, in particularly, to address collinearities that are very strongly evident in gene expression studies.

An extension of the above approach begins to allow this by considering multiple orderings of the variables prior to generating DAGs, as follows. At each step of the iterative ordering of variables, we have a set of candidate variables for selection as next in the order, each with its current score. Rather than selecting just the variable with the highest score, we can consider several possible orderings using a set of variables with high scores, and/or sample from the variables according to some probabilities. This is easily implemented and underlies our analyses for the examples here. Specifically, we use an annealed approach that draws variables i according to probabilities proportional to $(s_i^O)^a$ for some annealing parameter $a > 0$, so introducing stochastic variation into the generation of the ordering. Repeatedly developing models this way leads to multiple candidate DAGs, each with their resulting posterior probabilities, and thus classes of high probability graphical models. In our experiments with several datasets, this strategy indeed identifies graphs of higher posterior probability than the direct deterministic search (the latter corresponding to $a \to \infty$), though too small a value of a leads to the generation of models that are more widely dispersed in graph space and can have much lower probability than those found by deterministic search. Based on initial experiments, we use $a = 25$ in the example analyses.

5 Breast Cancer Gene Expression Analyses

We have developed the analysis on expression data from 158 breast cancer samples arising in our studies of molecular phenotyping for clinical prediction [14–16]. Snapshots of this analysis here focus on how aspects of associations exhibited in generated graphs relate to known biology and how other aspects may suggest biological investigation. Our analysis fits the model to the full set of 12,558 probe sets (after deleting 67 controls) on the Affymetrix U95aV2 microarrays; we use log2, full quantile normalized transforms of the Affymetrix MAS5.0 signal measure of expression [16]. The analysis adopts hyper-parameter $\beta = 1/(p-1)$, a prior that is consistent with sparsity and constrains proliferation of edges in the resulting graphs. The prior mean of $k_i = \#cp(i)$ is then 1, with prior probability of about 0.3 on each of $k_i = 0, 1$. Using the annealing parameter $a = 25$, the analysis was run as described, and a total of 150 graphs saved. The summary examples draw on the highest scoring graph, an approximate local posterior mode in graph space, together with some aspects of additional high scoring graphs.

5.1 Estrogen Receptor Pathway Genes

A proof-of-principle exploration focuses on small pieces of the resulting graphs related to the estrogen receptor (ER) gene. ER and the estrogen pathway play key roles in breast cancer and the evolution and behavior of tumours [17–19]. ER is a transcription factor that directly regulates a variety of genes in an estrogen-dependent manner [20], and so this component of the network is of key interest. One such ER target is the TFF1 gene, shown in past work to be a direct target of ER with an estrogen response element (ERE) within the critical promoter sequences [21]. As a starting point in the analysis of the network, we extracted the subgraph on a set of genes known to lie in the ER pathway and relate to TFF1, and identified all links between these genes in the top graph generated as well as ten others (Figure 1). Most of the genes here have additional neighbors, not shown. Two key transcription factors are

FOXA1 (HNF3a) and GATA3. FOXA1 has been shown to play a direct role in the transcription of the TFF1 gene [22]; the TFF1 promoter contains a binding site for FOXA1 as well as multiple GATA sites. Also, though we are not aware of work showing a direct role for GATA3 in the transcription of TFF1, the related GATA6 protein is so implicated and this is dependent on the GATA elements [23]. The graphs also give a very clear indication of an intermediary role for TFF3, another member of the intestinal trefoil factor family and close cousin of TFF1. Hence small paths in the graph that link ER with TFF1, based on predictive/regression based dependencies of expression of these genes, reveal a set of activities that directly participate in the control of TFF1 expression as well as insights into other potential interactions.

This sub-graph has additional features linked to known functional interactions. FOXA1 links directly to androgen receptor (AR), which is known to regulate ER expression [24]. Also, a direct link to FOXA1 is the c-MAF oncogene which also links to FOXF1. MAF is a transcription factor that inhibits MYB activity [25], and MYB is a key link from ER to TFF1. FOXF1 is a member of the forkhead family of transcription factors, and several other forkhead family members do interact with ER [26]. Our analysis directly implicates FOXF1 and so suggests a broader role for the family in interactions with the ER pathways, promoting an interest in further biological study of FOXF1. A further key gene, IGF1R, is seen to directly link to ER and also MYB. MYB is known to play a role in activating IGF1R [27]; also, IGF1, the ligand for IGF1R, has been shown to be regulated by AR [24]. So we reveal a series of functional interactions involving the genes that form the paths between ER and TFF1 as identified by patterns of association in expression, and also clues about other, incompletely understood or as yet undefined biological interconnections. Other interpretable features include additional genes that are known to be regulated by, or co-regulated with ER (such as LIV-1 and BCL-2), and additional clues about potential biological interconnections with genes not shown. Even at this first cut with one initial candidate model, the potential use of sparse statistical graphs in reflecting known biology in very high-dimensional data as well as in generating insights and clues for further study is clear.

Much more information can be generated by successfully exploring genes in neighborhoods of genes of interest. One positive point of note is that the transcription factors FOXA1 and GATA3 here have many more first-order neighbors than other genes displayed. This is natural in that biologically key transcription factors are expected to have a sustained influence on cascades of downstream genes, and the graph reflects this. The broader understanding of the roles of key transcription factors may be enhanced by such studies through the generation of investigations of sets of genes in such neighborhoods.

On a technical point, the Affymetrix data includes multiple probe sets for some genes, including ER (ESR1, HG3125-HT3301), MYB (U22376, MYBa, MYBe, MYBf), AR (AR, ARa), c-MAF (MAF, MAFa), TFF3 (TFF3, TFF3b), XBP

(XBP1, XBP1a) and IGF1R (IGF1R, IGF1Ra) in this sub-network. The tight linking of multiple probe sets for one gene in the graph is positive, and a facility to "collapse" together the corresponding node would underscore the above discussion. This feature is also useful in identifying probe sets (such as other probe sets for MYB not shown) that do not tie-in so tightly to their sibling probe sets, perhaps reflecting on the issue of probe design and selection.

5.2 RB/E2F Pathway Exploration

The Rb-E2F pathway is the key regulatory process governing the transition of cells from a quiescent state to a growing state [28,29]. The E2F transcription factors regulate a large group of genes involved in DNA replication, mitosis, and cell cycle progression. The Rb protein controls the activity of the pathway by binding to and regulating the E2F factors. As an example in using the graphical networks to extend the understanding of the Rb pathway, we identify a sub-graph that contains genes known to function in the Rb-E2F pathway; see Figures 2 and 3. Figure 2 depicts the general outline of the Rb-E2F pathway with a number of genes that are known to function in the pathway. This includes a selected group of known targets for the E2F transcription factors; these targets are grouped according to roles in DNA replication, mitosis, cell cycle progression, and the apoptopic response. Genes used for the query of the graph are indicated in green, with those appearing in the sub-graph represented as solid green symbols. The sub-graph generates other genes as linkers between query genes. This provides a first step to the discovery of other genes potentially linked to the pathway.

Of 48 genes used to search the network, all but 11 appeared in proximity in the sub-network (green genes in Figure 2). The search was limited to those genes that were connected by no more than two edges in the top scoring graph, thus restricting the number of genes to those most closely related in expression to the search group. That so many genes connect in the resulting sub-network argues strongly that their association relates to the functional context. Indeed, similar searches on random collections of genes yields highly disconnected subgraphs with far fewer edges. It is also clear that many of the genes discovered in the Rb-E2F sub-network are in fact related to Rb-E2F pathway function. Of the 33 genes discovered as "linkers" between genes in the query, 13 (in red) can be readily be identified as functioning in the pathway based on a literature search. Thus the graphical models appear to provide information relevant to extending understanding of gene regulatory pathways. One key exploratory use of these models is clearly evident: the additional linker genes discovered that are not easily explained or already understood to play functional roles in the pathway are now candidates for further study based on being implicated in the function of the pathway through predictive association in the graphical model.

6 Additional Comments

The example represents a graphical model in $p = 12,558$ dimensions. We are unaware of previous work that attempts analysis in such high dimensions. It is very important to fit the entire distribution – that is, to consider all genes – even though substantive interest and evaluation must focus on small subsets. Otherwise, we may find associations induced through variables not included in the analysis, with potential to mislead. In even moderate dimensions, sparsity is critical in aiding implementation, since sparser graphs are much easier to generate and explore, and we again stress the relevance and criticality of sparsity from the biological point of view.

The two examples from analysis of breast cancer data begin to demonstrate the potential utility of such models in generating biological insights and gene candidates related to a specific pathway of interest. We note that, beyond exploring the analysis outputs based on model fits to a single data set, it will be of broader interest to develop analyses of multiple data sets and some of the future potential lies in exploring and contrasting structure of sub-graphs on specific sets of genes in a known pathway.

Current developments include foci on the key next steps statistically: developing stochastic search and model selection procedures to generate not just local posterior modes in model space but multiple, representative models and sets of high-posterior probability graphs as a result. Biological interpretation and insights require some appreciation of the limitations of one selected model, and generating an expanding set of plausible graphs, will aid in this. Further, in even very low dimensions high and complex patterns of collinearity among genes is the norm, and to deal properly with this requires the evaluation of many plausible models. The approach here does indeed explore multiple models with small, stochastic perturbations of the deterministic search algorithm, as a start in this direction. Our current research focuses on developing more advanced stochastic search methods with this goal.

It is tempting to imagine developments towards more extensive, confirmatory styles of analysis could in principle begin to build biological information into the prior distributions, especially in connection with known and expected pathway interactions, but we are some way from developing appropriate formalisms; at this point, we stress the use and development of exploratory analyses with the more "unbiased" priors described here as a tool to aid in structure discovery and elucidation. It is likely that this utility will be enhanced substantially by the development of methods to integrate with information systems that provide for similar exploration of biological/literature databases, so that the gene discovery process can iterate between statistical discovery and biological confirmation.

Acknowledgments

Research reported was partially supported by the National Science Foundation on grant DMS-0102227 and 0112340, through SAMSI under DMS-0112069, and by the NIH. Graphical displays are based on the AT&T *graphviz* software [30]. The authors would like to thank the editors and three anonymous reviewers whose comments helped substantially in the revisions.

References

[1] M. West, Bayesian factor regression models in the "large p, small n" paradigm, in: J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West (Eds.), Bayesian Statistics 7, Oxford University Press, 2003, pp. 723–732.

[2] X. Zhou, M. J. Kao, W. H. Wong, Transitive functional annotation by shortest path analysis of gene expression data, Proceedings of the National Academy of Sciences 99 (2002) 12783–12788.

[3] R. Hofmann, V. Tresp, Nonlinear Markov networks for continuous variables, in: M. I. Jordan, M. J. Kearns, S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference, MIT press, 1998, pp. 521–527.

[4] S. L. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.

[5] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, C. Kadie, Dependency networks for inference, collaborative filtering, and data visualization, Journal Of Machine Learning Research 1 (2000) 49–75.

[6] A. Roverato, G. Consonni, Compatible prior distributions for DAG models, DIMACS Technical Report 2002-17 (2002).

[7] D. J. Spiegelhalter, S. L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, Networks 20 (1990) 579–605.

[8] A. P. Dawid, Some matrix-variate distribution theory: Notational considerations and a Bayesian application, Biometrika 68 (1981) 265–274.

[9] P. Giudici, Learning in graphical Gaussian models, in: J. M. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), Bayesian Statistics 5, Oxford University Press, 1994, pp. 621–628.

[10] D. Geiger, D. Heckerman, Parameter priors for directed acyclic graphical models and the characterization of several probability distributions, Annals of Statistics 5 (2002) 1412–1440.

[11] A. P. Dawid, S. L. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable graphical models, The Annals of Statistics 3 (1993) 1272–1317.

[12] A. Zellner, An Introduction to Bayesian Inference in Econometrics, New York: Wiley, 1971.

[13] S. A. Anderson, D. Madigan, M. D. Perlman, A characterization of markov equivalence classes for acyclic digraphs, Annals of Statistics 25 (1997) 505–541.

[14] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, M. J.R., J. R. Nevins, Predicting the clinical status of human breast cancer using gene expression profiles, Proceedings of the National Academy of Sciences 98 (2001) 11462–11467.

[15] E. Huang, M. West, J. R. Nevins, Gene expression profiles and predicting clinical characteristics of breast cancer, Recent Progress in Hormone Research (2003) 55–73.

[16] E. Huang, S. Cheng, H. Dressman, J. Pittman, M.-H. Tsou, C.-F. Horng, A. Bild, E. Iversen, M. Liao, C.-M. Chen, M. West, J. Nevins, A. Huang, Gene expression predictors of breast cancer outcomes, Lancet 361 (2003) 1590–1596.

[17] J. A. Henry, S. Nicholson, J. R. Fandon, B. R. Westley, F. E. May, Measurement of oestrogen receptor mRNA levels in human breast tumours, J Breast Cancer 58 (1988) 600–605.

[18] W. A. Knight, R. B. Livingston, E. J. Gregory, W. L. McGuire, Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer, Cancer Res 37 (1977) 4669–4671.

[19] M. F. Pichon, P. Broet, H. Magdelenat, J. C. Delarue, F. Spyratos, J. P. Basuyau, S. Saez, A. Rallet, P. Courriere, R. Millon, B. Asselain, Prognostic value of steroid receptors after long term follow up of 2257 operable breast cancers, Br J Cancer 73 (1996) 1545–1551.

[20] F. E. May, B. R. Westley, Identification and characterization of estrogen-regulated RNAs in human breast cancer cells, J Biol Chem 263 (1988) 12901–12908.

[21] T. Barkhem, L. A. Haldosen, J. A. Gustafsson, S. Nilsson, Gene expression in HepG2 cells: Complex regulation through crosstalk between the estrogen receptor alpha, an estrogen-response element, and the activator protein 1 response element, Mol Pharmacol 61 (2002) 1273–1283.

[22] S. Beck, P. Sommer, E. Do Santos Silva, N. Blin, P. Gott, Hepatocyte nuclear factor 3 (winged helix domain) activates trefoil factor gene TFF1 through a binding motif adjacent to the TATA box, DNA Cell Biol 18 (1999) 157–164.

[23] E. Al-azzeh, P. Fegert, N. Blin, P. Gott, Transcription factor GATA-6 activates expression of gastroprotective trefoil genes TFF1 and TFF2, Biochem Biophys Acta 1490 (2000) 324–332.

[24] L. Sahlin, G. Norstedt, H. Eriksson, Androgen regulation of the insulin-like growth factorI and the estrogen receptor in rat uterus and liver, J Steroid Biochem Mol Biol 51 (1994) 57–66.

[25] S. P. Hegde, A. Kumar, C. Kurschner, L. H. Shapiro, c-Maf interacts with c-Myb to regulate transcription of an early myeloid gene during differentiation, Mol Cell Biol 18 (1998) 2729–2737.

[26] E. R. Schuur, A. V. Loktev, M. Sharma, Z. Sun, A. A. Roth, R. J. Weigel, Ligand-dependent interaction of estrogen receptor alpha with members of the forkhead transcription factor family, J Biol Chem 276 (2001) 33554–33560.

[27] K. Reiss, A. Ferber, S. Travali, P. Porcu, P. D. Phillips, R. Baserga, The protooncogene c-Myb increases the expression of insulin-like growth factor 1 and insulin-like growth factor 1 receptor messenger RNAs by a transcriptional mechanism, Cancer Res 51 (1991) 5997–6000.

[28] J. R. Nevins, Towards an understanding of the functional complexity of the E2F and retinoblastoma families, Cell Growth Differ. 9 (1998) 585–593.

[29] N. Dyson, The regulation of E2F by pRB family proteins, Genes Dev. 12 (1998) 2245–2262.

[30] GraphViz, Open source graph drawing software, AT&T Research Labs., http://www.research.att.com/sw/tools/graphviz/.

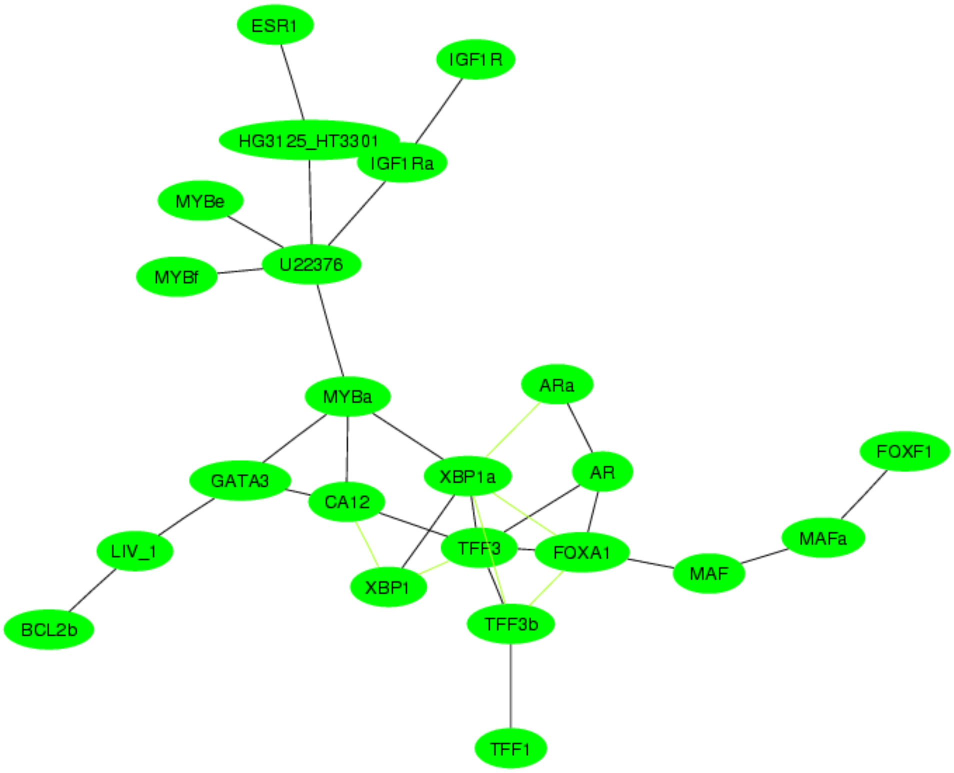

Fig. 1. Sub-graphs of breast cancer based gene expression graphs (on 12,558 genes) involving ER and TFF1. Some genes have multiple sets of oligonucleotide sequences on the microarray, and hence the appearance of multiples of some genes: estrogen receptor itself (ESR1, HG3125-HT3301), MYB (U22376, MYBa, MYBe, MYBf), AR (AR, ARa), c-MAF (MAF, MAFa), TFF3 (TFF3, TFF3b), XBP (XBP1, XBP1a) and IGF1R (IGF1R, IGF1Ra). The most probable graph generated has edges indicated in black. The additional green edges are those appearing in at least 5 of the top 10 graphs generated.

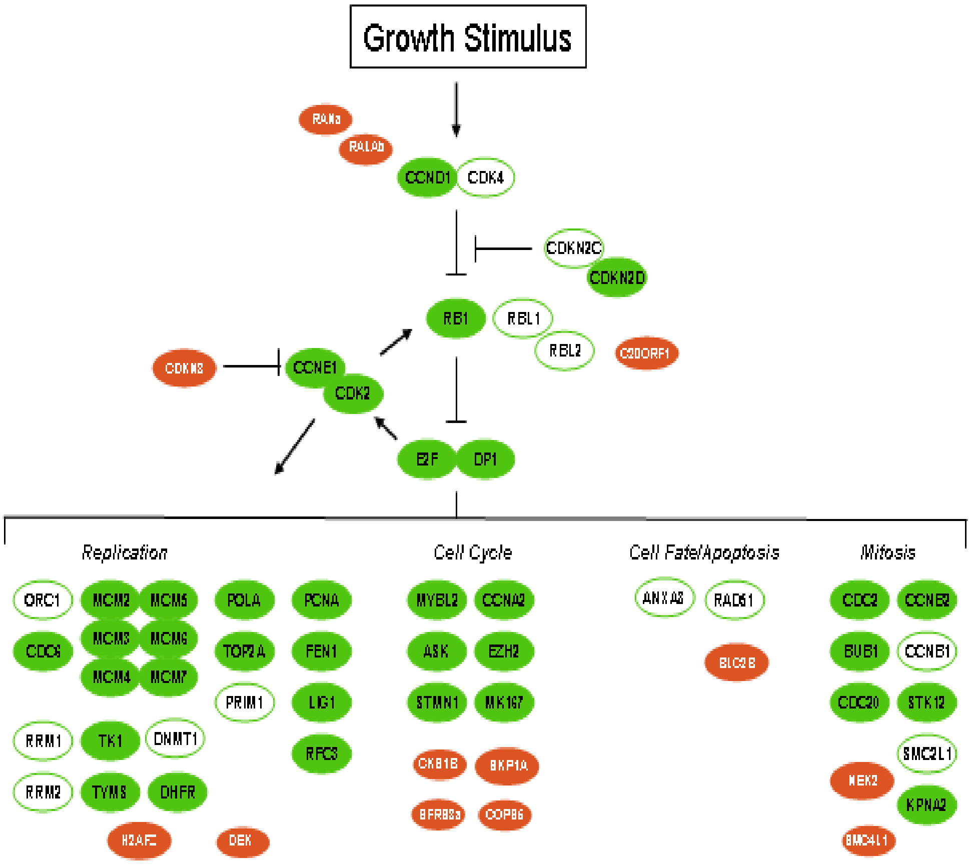

Fig. 2. Genes related to the Rb-E2F pathway. The initial set of query genes was based on genes known to function in the Rb-E2F pathway and they are grouped by functional role. Those appearing directly connected in the sub-graph appear as solid green symbols, the others as open green symbols. The discovered linker genes – that lie between pairs of queries in the sub-graph – appear as red, and are placed here according to known or suspected function in the pathway.

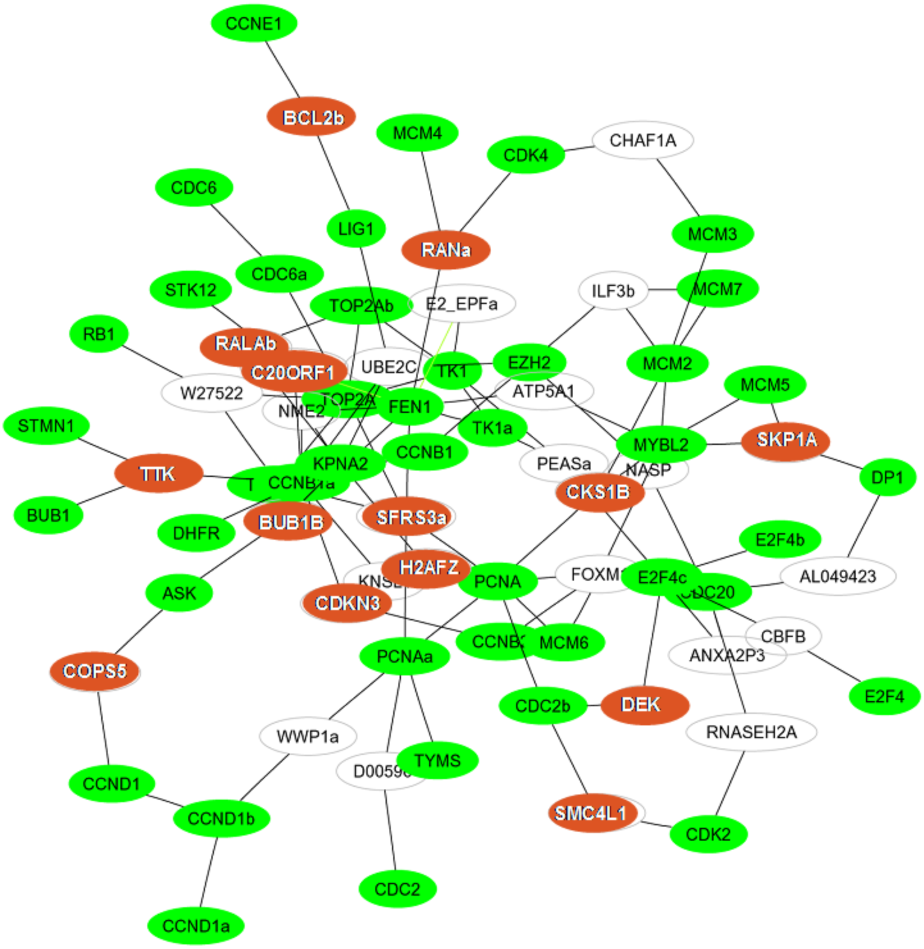

Fig. 3. The sub-graph containing Rb-E2F pathway query (green) and linker (red and uncolored) genes as described in Figure 2, arising from the exploring the first-order neighbors of the set of query genes in the most highly scored, 12,588 node gene expression graphical analysis of the breast cancer data.