BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–11

TREEGL: Reverse Engineering Tree-Evolving Gene Networks Underlying Developing Biological Lineages

Ankur P. Parikh [†,1], Wei Wu [†,2], Ross E. Curtis [3,4] and Eric P. Xing [1,3,4 *]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[2]Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213
[3]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213
[4]Joint Carnegie Mellon University-University of Pittsburgh PhD Program in Computational Biology, Pittsburgh, PA, 15213
[†] Authors contributed equally

ABSTRACT

Motivation: Estimating gene regulatory networks over biological lineages is central to a deeper understanding of how cells evolve during development and differentiation. However, one challenge in estimating such evolving networks is that their host cells not only contiguously evolve, but also branch over time. For example, a stem cell evolves into two more specialized daughter cells at each division, forming a tree of networks. Another example is in a laboratory setting: a biologist may apply several different drugs individually to malignant cancer cells to analyze the effects of each drug on the cells; the cells treated by one drug may not be intrinsically similar to those treated by another, but rather to the malignant cancer cells they were derived from.

Results: We propose a novel algorithm, *Treegl*, an ℓ_1 plus total variation penalized linear regression method, to effectively estimate multiple gene networks corresponding to cell types related by a tree-genealogy, based on only a few samples from each cell type. *Treegl* takes advantage of the similarity between related networks along the biological lineage, while at the same time exposing sharp differences between the networks. We demonstrate that our algorithm performs significantly better than existing methods via simulation. Furthermore we explore an application to a breast cancer dataset, and show that our algorithm is able to produce biologically valid results that provide insight into the progression and reversion of breast cancer cells.

Availability: Software will be available at http://www.sailing.cs.cmu.edu/

Contact: epxing@cs.cmu.edu

1 INTRODUCTION

A major challenge in systems biology is to quantitatively understand and model the topological and functional properties of cellular networks, such as the transcriptional regulatory circuitries and signal transduction pathways that control cell behavior in complex biological processes. In complex organisms, biological processes such as differentiation and development are often controlled by a large number of molecules that exchange information in a spatial-temporally specific and context-dependent manner. These cellular networks are inevitably changing to take on different functions and reacting to changing environments. This necessitates studying different networks for each condition, such as each different developmental stage, tissue subtype, and cell lineage.

Most existing techniques for reconstructing molecular networks based on high-throughput data ignore the intricate dependencies between networks of closely related biological subjects.

For example, when studying cancer development, it is common to infer gene networks based on microarray data from different cancer specimens or cell lines *separately* and *independently*, despite that these biomaterials are usually collected over a contiguous disease progression course. As we discuss in detail, such an "isolationist" strategy can compromise both the statistical power and biological insight of the inferred networks. In this paper, we present a new methodology called *Treegl*, which adopts a statistically more powerful and biologically more natural "connectionist" principle. *Treegl* reconstructs gene networks in related biological subjects via an *inter-dependent* approach such that the inferred networks directly embody and exploit the relationships of the biological subjects they represent. As a result, this reveals deeper insight on how the structure, function, and behavior of such networks evolve during evolution, differentiation, and environmental perturbation.

To better understand our rationale, take the analysis of stem cell differentiation as an example. It is well known that all organ- and tissue-specific cells in a multicellular organism are differentiated from a stem cell, following a well-known genealogy (Figure 1). To date, gene networks from many of these organs and tissues have been derived using a variety of computational or experimental technologies (Basso *et al.*, 2005; Li, 2004; Hyatt *et al.*, 2006). However, knowledge about the cell lineage has rarely been utilized in constructing these networks. For example, according to the genealogy in Figure 1, the platelets are more closely related to the red blood cells than to the lymphoblasts. The gene networks present in the platelets and red blood cells are thus expected to be more similar; microarray data from red blood cells should reflect the topology of a platelet's network to a greater extent than that of a lymphoblast network.

Is it therefore legitimate and possible to use the red blood cell's microarray in addition to the platelet's microarray to infer the platelet's network? And, if yes, how? Essentially, what one needs to handle is a *network of networks*. In this paper, we focus on the class of tree-shaped biological genealogies. This class of genealogies can be naturally found in crop and animal breeding, species evolution, cell-line lineage construction, and carcinogenesis.

*To whom correspondence should be addressed.

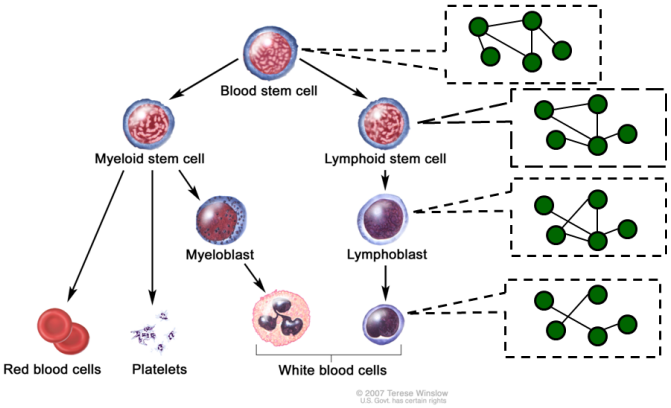

Fig. 1. A tree-evolving network in the blood stem cell genealogy (*http://www.siteman.wustl.edu/CancerDetails.aspx?id=661&xml=CDR257990.xml*). At the root is the blood stem cell. Over time it differentiates into more specialized cells along the genealogy, eventually becoming red blood cells, platelets, or white blood cells. Cell types closer together in the genealogy (e.g., lymphoblasts and white blood cells), are expected to have more similar gene regulatory networks than those far apart in the genealogy (e.g., red blood cells and white blood cells).

1.1 Related work

There has been a lot of previous work on reverse engineering gene networks. However, most of this work revolves around estimating a static network, losing the dynamic information that we seek to explore and exploit. For example, Friedman *et al.* 2000 proposed using Bayesian networks to reverse engineer gene networks. However, their method assumed all the measurements of gene expression from the network in question were independent and identically distributed (*i.i.d.*) from the same distribution, and introduced extra variables to try to capture certain stationary (rather than time-evolving) time dependence. Furthermore, their algorithm was not scalable to the high dimensional problems that we are considering. Margolin *et al.* 2006 proposed an information theoretic approach that has good statistical properties, but limits the network structure to having neglible loops. Yeung *et al.* 2002 proposed using a singular value decomposition. Like Friedman *et al.* (2000), these methods also assumed the data were *i.i.d.* from an invariant network.

Recently, researchers have begun tackling the time-varying case, building off sparse regression techniques, like the lasso (Tibshirani, 1996). Lozano *et al.* proposed an approach that uses the group lasso and the notion of Granger causality to estimate causality among variables instead of estimating the entire sequence of networks (Lozano *et al.*, 2009). Bonneau et al.2006 propose using the kinetic equation in conjunction with the lasso to account for time series data (but also learn only one network). Ahmed and Xing created TESLA (Ahmed and Xing, 2009), and Song, Kolar and Xing proposed KELLER (Song *et al.*, 2009a), to estimate a chain of evolving networks over time. Song *et al.* also proposed time varying dynamic Bayesian networks (Song *et al.*, 2009b).

However, all these methods estimate networks that evolve as a chain of graphs over time, not a genealogy, which hinders them from being naturally applied to many of the common biological applications mentioned earlier.

1.2 Our contribution

In this work, we move beyond the static and time-varying assumptions, and focus on the more general case of tree-evolving genealogies that we believe are more natural for the biological phenomena that we seek to explore. We propose an algorithm called *Tree-smoothed graphical lasso* (*Treegl*), that can effectively and jointly recover evolving regulatory networks present in multiple cell-types related by a tree genealogy.

Our approach takes advantages of the similarities of networks nearby in the genealogy, but can also reveal sharp differences. Moreover, by building on the method of neighborhood selection via the lasso (Meinshausen and Bühlmann, 2006), our approach works well even when the number of genes is much larger than the number of samples.

We were motivated by the many applications discussed above in the development of *Treegl*. However, in this paper, we focus on applying *Treegl* to study the progression and reversion of breast cancer cells in 3-dimensional organotypic cultures (Weaver *et al.*, 1997; Liu *et al.*, 2004; Itoh *et al.*, 2007). The cell-line in question begins as nonmalignant, organized, and nontumorigenic cells that progress to apolar, disorganized, and tumorigenic cancer cells. Several different drugs are applied and the genealogy then branches to different reverted cells with partially polarized structures. Although our dataset is small, we are able to show that we obtain biologically valid and intriguing results through our method.

2 METHODS

2.1 Probabilistic Representation of Gene Networks

Consider the problem of modeling N different, but independent (we will consider dependency in the next subsection), gene regulatory networks, each corresponding to a unique cell type (say, type n) from a cell bank $\mathcal{B}$ where $|\mathcal{B}| = N$, with S_n *i.i.d.* microarray measurements of all genes in cell type n, and consisting of the same set of p genes across all cell types. Without loss of generality, a gene network can be represented by a probabilistic graphical model, such as a Markov random field (MRF) if the gene states are taken as discrete (Segal *et al.*, 2003), or a Gaussian graphical model (GGM) if the gene states are set to the continuous measurements of the microarray signal (Dobra *et al.*, 2004), or a Bayesian network (Friedman *et al.*, 2000). In this paper, we use cell-type specific undirected Gaussian graphical models to model the gene networks, but the general principle of our method can be extended to discrete Markov random fields as well.

Let $\mathcal{G}^{(n)} = (\mathcal{V}^{(n)}, \mathcal{E}^{(n)})$ represent a network in cell type n, of which $\mathcal{V}^{(n)}$ denotes the set of genes, and $\mathcal{E}^{(n)}$ denotes the set of edges over vertices. An edge $(u, v) \in \mathcal{E}^{(n)}$ can represent a relationship (e.g., influence or interaction) between genes u and v. Let $\mathbf{X}^{(n,s)} = (X_1^{(n,s)}, \ldots, X_p^{(n,s)})'$, where $n \in \mathcal{N}$, $s \in \{1, \ldots, S_n\}$, and $p = |\mathcal{V}|$, be a random vector of nodal states that are real valued and standardized, such that each dimension has mean 0 and variance 1. We assume that $\mathbf{X}^{(n)}$ follows a multivariate Gaussian distribution with mean 0 and covariance matrix $\Sigma^{(n)}$, so that the conditional independence relationships among the genes can be encoded as a Gaussian graphical model. It is a well known fact that for GGMs, edges in the graph correspond to non-zero elements in the inverse covariance matrix (known as the precision matrix), which we denote by $\mathbf{\Omega}^{(n)} := (\omega_{uv}^{(n)})_{u,v \in [p]}$. Thus, estimating the

graph structure is equivalent to selecting the non-zero elements of the precision matrix.

As commonly done, instead of directly estimating the precision matrix elements $\omega_{uv}^{(n)}$, we estimate the partial correlation coefficients $\boldsymbol{\rho}^{(n)}$, where $\rho_{uv}^{(n)}$ is the correlation between gene u and gene v conditioned on the values of all the other genes. Partial correlation coefficients are related to the precision matrix elements by Eq. 1.

$$\rho_{uv}^{(n)} = -\frac{\omega_{uv}^{(n)}}{\sqrt{\omega_{uu}^{(n)}\omega_{vv}^{(n)}}}. \tag{1}$$

As shown in Eq. 1, $\rho_{uv}^{(n)}$ is zero if and only if $\omega_{uv}^{(n)}$ is zero. Therefore, in terms of network structure estimation, the network resultant from the non-zero $\rho_{uv}^{(n)}$ is equivalent to that from the nonzero $\omega_{uv}^{(n)}$. Furthermore, the partial correlation is quite intuitive in the sense that a high positive value of $\rho_{uv}^{(n)}$ indicates that the genes u and v are strongly positively correlated (conditioned on the other genes), a low negative value indicates the genes are strongly negatively correlated (conditioned on the other genes), and $\rho_{uv}^{(n)} = 0$ for all $(u, v) \notin \mathcal{E}^{(n)}$. As a result, we simply consider estimating the partial correlation coefficients and designate these as the edge values in $\mathcal{G}^{(n)}$:

$$\mathcal{E}^{(n)} = \{\rho_{uv}^{(n)} : |\rho_{uv}^{(n)}| > 0\}. \tag{2}$$

2.2 Neighborhood Selection

A number of recent papers have studied how to estimate this model from data that are assumed to be *i.i.d.* samples from the model, and the asymptotic guarantee of the estimator (Wainwright *et al.*, 2007; Bresler *et al.*, 2008). In particular, an efficient neighborhood selection algorithm (Meinshausen and Bühlmann, 2006) based on ℓ_1-norm regularized regression has been proven effective (often called neighborhood selection). In this approach, the neighborhood of each gene u is estimated independently using a penalized linear regression with a lasso-style (i.e., ℓ_1-norm) regularization over edge weights. The regression goes around every gene in the network, leading to completion of a network. In every neighbor estimation step, gene u is treated as a response variable, all the other genes are the covariates, and the weights are the correlations between the other genes and u. More formally, let $\mathbf{X}_{\setminus u}$ indicate the $p-1$ vector of the values of all genes except u. Similarly, $\boldsymbol{\theta}_{\setminus u} := \{\theta_{uv} : v \in \mathcal{V} \setminus u\}$. Using a well known result (Lauritzen, 1996) that the partial correlation coefficients can be related to the following regression model:

$$X_u^{(n,s)} = \sum_{v \neq u} X_v^{(n,s)} \theta_{uv}^{(n)} + \epsilon_u^{(n,s)}, \quad u \in [p], \tag{3}$$

where $\epsilon_u^{(n,s)}$ is uncorrelated with $\mathbf{X}_{\setminus u}^{(n,s)}$ if and only if

$$\theta_{uv}^{(n)} = -\frac{\omega_{uv}^{(n)}}{\omega_{uu}^{(n)}} = \rho_{uv}^{(n)} \sqrt{\frac{\omega_{vv}^{(n)}}{\omega_{uu}^{(n)}}}. \tag{4}$$

Some algebra gives that

$$\rho_{uv}^{(n)} = \text{sign}(\theta_{uv}^{(n)}) \sqrt{\theta_{uv}^{(n)} \theta_{vu}^{(n)}}. \tag{5}$$

The above equations basically indicate that we can solve for the regression coefficients $\boldsymbol{\theta}_{\setminus u}$ using a linear regression, where the response variable corresponds to X_u and the covariates correspond to $\mathbf{X}_{\setminus u}$. The corresponding partial correlation coefficients can be recovered using Eq. 5. An ℓ_1 penalty is applied to encourage a sparse solution, as in the lasso (Tibshirani, 1996).

This surprisingly simple method, when applied over *i.i.d.* nodal samples (e.g., *i.i.d.* microarray measurements), has very strong theoretical guarantees about recovering the correct network structure. It has been shown that under certain variable conditions it is possible to obtain an estimator of the edge set $\mathcal{E}$ that achieves a property known as *sparsistency* (Meinshausen and Bühlmann, 2006; Wainwright *et al.*, 2007), which refers to the case where a consistent estimator of $\mathcal{E}$, i.e., the network structure, can be attained when the true degree (i.e., number of neighbors) of each node is much smaller than the size of the graph p (even when the sample size is significantly smaller than the number of genes).

Unfortunately, in the case of the tree-evolving network concerned in this paper, we have to deal with a much harder problem since our samples are no longer *i.i.d.*, and our networks are no longer independent of each other. For this purpose, we need to extend the basic neighborhood selection lasso algorithm as shown in the following subsections.

2.3 Tree-Evolving Gene Networks Over Biological Lineages

We are interested in reconstructing *a set of networks* $\mathcal{G}^{(1)}, ..., \mathcal{G}^{(N)}$ that are not independent of each other, but are related by a genealogy over their respective host cell-types, thereby constituting a tree evolving network. Formally, given a genealogy over members of a cell bank $\mathcal{B}$, we introduce an ordering over networks $\mathcal{G}^{(1)}, ..., \mathcal{G}^{(N)}$ encoded by the following inheritance relationship: for each cell type $n \in \mathcal{B}$, let $\pi(n)$ be the parent of type n in the tree, thus $\mathcal{G}^{(n)}$ is a *descendant* of $\mathcal{G}^{(\pi(n))}$. For a pair of networks identified by the genealogy, we assume that their topology should be *similar* while allowing for differences. For example, consider again Figure 1. In this case, π(blood stem cell) = NULL, π(lymphoid stem cell) = blood stem cell, π(lymphoblast) = lypmhoid stem cell, etc. Note that this framework is flexible and allows for various types of trees since each parent can have a different number of children. We assume without loss of generality that $\mathcal{G}^{(1)}$ is the root of the tree.

Based on the GGM representation of gene networks described in the previous subsection, we have a set of GGMs whose edges (partial correlation coefficients) $\boldsymbol{\rho}^{(n)}, \forall n$ are evolving across the genealogy. Since the partial correlations are functions of the conditions rather than constants such a model is an instance of a *varying-coefficient model* (Fan and Yao, 2005). Varying-coefficient models were popularized in the work of (Cleveland and Grosse, 1991) and (Hastie and Tibshirani, 1993), and have been applied to a variety of domains to model and predict time- or space- varying response to multidimensional inputs. In our case, we are particularly interested in a certain type of parameter change: the change between zero and non-zero values between $\boldsymbol{\rho}^{(n)}$ and $\boldsymbol{\rho}^{(\pi(n))}$, also known as the *structural change* of the model.

The tree evolving networks described above are effective for modeling a plethora of biological processes such as the growth and reversion of cancer. A biologist may apply several treatments to a malignant cancer cell and would like to analyze the effects of the treatments on the regulatory network. The tree structure naturally expresses the dependence of the treated cells on the malignant cell

without forcing the two treated cells to be identical. We explore this application in more detail later in the paper.

2.4 Estimating Tree-Evolving Networks

When the network is tree-evolving, our goal is to learn the *structure* of a tree-varying GGM, which is a special case of the general varying-coefficient varying-structure (VCVS) model studied in (Kolar *et al.*, 2009). This formulation allows us to formally encode the topology of the network into the parameters $\boldsymbol{\rho}^{(n)}$ of the model; for example, the absence of an edge between nodes u and v in cell type n, corresponds to the partial correlation coefficient $\rho_{uv}^{(n)} = 0$.

Thus, in our formulation, recovering the structure of the N gene regulatory networks in the cell genealogy can be done by estimating $\boldsymbol{\rho}^{(n)}$ for each $1 \leq n \leq N$. [1] Our goal is to capture the sharp differences (i.e., edge re-rewiring), rather than small correlation changes, in the tree evolving network. As a result, we concentrate on recovering the correct edge set $\mathcal{E}^{(n)}$ rather than on the exact values of $\boldsymbol{\rho}^{(n)}$, although these are attainable as a side product of our algorithm.

In line with this goal, we make three assumptions:

- Sparsity: Most of the $\rho_{uv}^{(n)}$ are zero, leading to graphs with few edges.
- Sparsity of change: The edge set $\mathcal{E}^{(n)}$ is similar to that of its parent $\mathcal{E}^{(\pi(n))}$.
- Sharpness of change: There do exist a few key differences between $\mathcal{E}^{(n)}$ and $\mathcal{E}^{(\pi(n))}$ that must be captured.

These assumptions hold in a wide variety of biological applications. Sparsity is usually well justified. For example, a transcription factor controls (and is controlled by) only a few genes under specific conditions (Davidson, 2001). A sparsity bias can effectively prevent estimating all elements in $\boldsymbol{\rho}^{(n)}$ to be non-zero, which leads to a meaningless complete graph. Similarly, in many biological processes the gene regulatory network in the parent cell type and the one in the child often contain only a few, but sharp differences. For example, if the parent network is a malignant cancer cell and the child networks are treated cancer cells with various drugs, we expect that the treated and cancer cells should have largely similar networks due to close developmental relationship. However, the genes that are affected by the drug should behave dramatically differently, causing a few large changes in the regulatory networks.

It is important to reiterate here that estimating networks for each cell type separately and independently is either invalid or extremely error-prone, because in common laboratory conditions only a few measurements of the gene expression are obtained, leading to either degeneracy of the likelihood function or high variance in the estimator. We overcome this problem by enabling information sharing across different cell types through a joint estimation of all networks under a *single* loss function, as opposed to a loss function defined on each individual network.

To estimate $\boldsymbol{\rho}^{(1)}, \ldots, \boldsymbol{\rho}^{(n)}$ jointly, we adopt the neighborhood selection idea described previously, and additionally penalize the difference between the neighborhoods of adjacent cell types in the genealogy. More specifically, to recover the neighborhood of gene u for all cell types jointly, we propose the following convex optimization problem for estimating tree evolving networks.

$$\hat{\boldsymbol{\theta}}_{\setminus u}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}_{\setminus u}^{(n)} = \underset{\boldsymbol{\theta}_{\setminus u}^{(1)}, \ldots, \boldsymbol{\theta}_{\setminus u}^{(n)}}{\arg\min} \left(\sum_{n=1}^{N} \sum_{s=1}^{S_n} \left(x_u^{(n,s)} - \boldsymbol{\theta}_{\setminus u}^{(n)} \mathbf{x}_{\setminus u}^{(n,s)} \right)^2 + \lambda_1 \sum_{n=1}^{N} \|\boldsymbol{\theta}_{\setminus u}^{(n)}\|_1 + \lambda_2 \sum_{n=2}^{N} \|\boldsymbol{\theta}_{\setminus u}^{(n)} - \boldsymbol{\theta}_{\setminus u}^{(\pi(n))}\|_1 \right). \quad (6)$$

In Eq. 6, $x_u^{(n,s)}$ refers to the realization of variable $X_u^{(n,s)}$. The ℓ_1 penalty associated with λ_1 enforces sparsity by setting most of the edge weights to 0 as shown in (Tibshirani, 1996). The total variation (TV) penalty associated with λ_2 enforces sparsity of difference and encourages most of the elements of $\boldsymbol{\theta}_{\setminus u}^{(n)}$ to be identical to those of $\boldsymbol{\theta}_{\setminus u}^{(\pi(n))}$ along the genealogy. However, since the ℓ_1 instead of the ℓ_2 penalty is used, outliers are not strongly penalized, allowing for large differences for a small set of edges. This allows us to have a large amount of information sharing among samples from related regulatory networks, while still allowing sharp differences to capture key changes as the network evolves.

One complication that results from the above approach is that since each neighborhood is estimated independently and because the regularization encourages some of the coefficients to be zero, the sign of $\hat{\theta}_{uv}^{(n)}$ is not guaranteed to equal the sign of $\hat{\theta}_{vu}^{(n)}$ for finite sample sizes. This makes directly using Eq. 5 to estimate the partial correlation coefficients difficult. One common way to address is "max" symmetrization, which is defined below.

$$\hat{\theta}_{uv}^{sym,(n)} = \begin{cases} \hat{\theta}_{uv}^{(n)} & : |\hat{\theta}_{uv}^{(n)}| \geq |\hat{\theta}_{vu}^{(n)}| \\ \hat{\theta}_{vu}^{(n)} & : |\hat{\theta}_{uv}^{(n)}| < |\hat{\theta}_{vu}^{(n)}| \end{cases} \quad (7)$$

We can now define our estimate of the partial correlation coefficients using Eq. 5 [2].

$$\hat{\rho}_{uv}^{(n)} = \text{sign}(\hat{\theta}_{uv}^{sym,(n)}) \sqrt{\hat{\theta}_{uv}^{sym,(n)} \hat{\theta}_{vu}^{sym,(n)}}. \quad (8)$$

The estimated edge set is then defined as:

$$\hat{\mathcal{E}}^{(n)} = \{(u,v) : |\hat{\rho}_{uv}^{(n)}| > 0\}. \quad (9)$$

The total variation penalty makes this algorithm significantly different from KELLER (Song *et al.*, 2009a). KELLER uses kernel reweighting to recover smoothly evolving networks where the correlations between genes are changing gradually over time. However, in both the stem cell evolution and breast cancer progression-reversion problems that motivate us, the networks are evolving sharply at some points while remaining almost constant in others. For example, different microarray measurements taken from a blood stem cell renewing itself while remaining in the undifferentiated state are expected to exhibit almost the same correlations among the genes. However, once the blood stem cell evolves into a myeloid or lymphoid stem cell as shown in Fig. 1, we expect there to be sharp changes in the regulatory network reflecting

[1] Note that this is technically not the pairwise potential function in a GGM

[2] Note that the symmetrization may not make this a good estimate of the magnitude of $\rho_{uv}^{(n)}$, but it is an accurate estimate of whether or not $\rho_{uv}^{(n)}$ is positive, negative, or zero, which is all we need to recover the network structure.

the new function of the more specialized cell. This sudden change can be effectively captured by the TV penalty in our algorithm but not by the kernel reweighting of KELLER. In this way, our algorithm is similar to that of TESLA (Ahmed and Xing, 2009) which also uses a TV penalty to estimate time evolving networks (a chain of graphs). However, our algorithm generalizes this idea to tree-evolving networks which are more suitable for investigating a wider range of biological processes. Algorithmically the genealogy-induced TV penalty defines more complex constraints on the model space than that of TESLA, where network structures should be inferred. It also uses a GGM approach, and thus involves a linear regression, instead of the binary MRF approach of TESLA, which involves a logistic regression. We believe that the GGM approach, which allows for continuous measurements, is more suitable for our breast cancer application, because the sample size is small.

2.5 Optimization

We employed the CVX solver (Grant *et al.*, 2008) provided in MATLAB to solve the underlying convex optimization problem for tree-evolving network estimation under our proposed model. At its core, CVX uses the SPDT3 solver (Toh *et al.*, 1999). SPDT3 is an interior point method for solving conic programming problems, where the constraints are convex cones, and the objective function is linear (plus the log-barrier terms for the constraints).

For larger scale problems, one can use the method proposed by Chen et al. 2010 that uses the accelerated gradient method.

3 SIMULATION RESULTS

To assess the performance of Treegl, we evaluated its performance on simulated microarray data with a known topology of the underlying tree-evolving network. Consider the following artificial tree evolving network with $N = 70$:

1. A graph A with 30 nodes, average degree 4, and max degree 6 is generated from a Gaussian Graphical Model. For the first 10 generations, i.e., $n = 1$ to 10, A remains unchanged. However, we assume that each of these generations correspond to a different cell type in the genealogy (for reasons that will be made clear later).
2. After $n = 10$, the graph branches into two child graphs, B and C. To generate each child graph, 25% of the edges are randomly deleted and the same number are randomly added. This represents a sharp, sparse change in the network. These child graphs stay unchanged for another 10 generations ($n = 11$ to 20 for B, $n = 21$ to 30 for C). Again, each generation indicates a different cell type.
3. B and C then branch further. 25% of the edges are randomly removed/added to generate graphs D and E from B, and F and G from C. The resulting graphs then stay constant for another 10 generations ($N = 31$ to 40 for D, $n = 41$ to 50 for E, $n = 51$ to 60 for F, and $N = 61$ to 70 for G).

Note that our algorithm does not know at which points the network structure changes. Our goal is to examine if it can detect the change-points as well as take advantage of the samples that come from cell types with identical structure between the change-points.

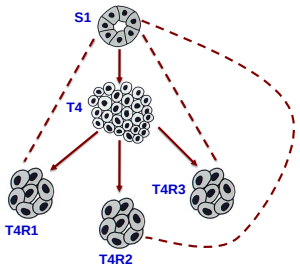

Fig. 3. Breast cancer genealogy. Solid arrows correspond to the genealogy. The dotted lines correspond to extra penalties between the T4R and S1 cells.

To evaluate Treegl, we plot a ROC curve showing the recall for different values of precision. $Precision = \frac{1}{N}\sum_{n=1}^{N}\frac{|\hat{\mathcal{E}}^{(n)}\cap\mathcal{E}^{(n)}|}{|\hat{\mathcal{E}}^{(n)}|}$, and $Recall = \frac{1}{N}\sum_{n=1}^{N}\frac{|\hat{\mathcal{E}}^{(n)}\cap\mathcal{E}^{(n)}|}{|\mathcal{E}^{(n)}|}$.

To produce the curve, cross validation is used to select λ_1 and λ_2. A threshold t is then varied from the smallest absolute edge weight to the largest absolute edge weight. An edge is included in the network if and only if it has an edge weight greater than t (in absolute value). We calculate precision/recall for a large number of values of t and produce the curve. To average different trials, we used binning, averaging points using a bin width of .05.

The results are shown in Figure 2 for two different sample sizes. Our method (in blue) performs favorably to estimating a single static network (green) or estimating each graph independently (red). It should be noted that our method can produce different graphs compared to the static method which only produces one. The independent method also produces different graphs but it performs very poorly.

4 AN APPLICATION TO BREAST CANCER DATA

We now demonstrate an application of our algorithm to the study of progression and reversion of breast cancer cells. Pioneered by Dr. Mina Bissell's research team, functional analysis of physiologically more realistic 3D culture models of breast cancer has yielded a wealth of insight into the mechanisms of cancer development (Petersen *et al.*, 1992). From tumor cells cultured in 3D matrices, it was found that microenvironmental factors and signaling inhibitors have a dramatic influence on the growth dynamics and malignancy of the cells (Weaver *et al.*, 1997; Itoh *et al.*, 2007). Further, tumorigenicity of breast cancer cells is tightly linked to the integrity of their acinar structures (Petersen *et al.*, 1992). However, except for a sketchy outline, little is known about how the cells interpret signaling cues from their surroundings and selectively regulate genes in a temporal-spatially specific manner.

Our goal is to investigate the gene regulatory networks of normal breast cells (S1 cells), malignant breast cancer cells (T4 cells), and nontumorigenic breast cancer cells reverted by different drugs (T4R cells). The exact tree-genealogy underlying these cell-type specific networks is shown in Figure 3: S1 cells with polarized acinar structures evolve into tumorigenic T4 cells which form disorganized apolar colonies, and then 3 drugs are applied individually to T4 cells and different reverted cells (T4R) with organized structures which resemble S1 cells are produced.

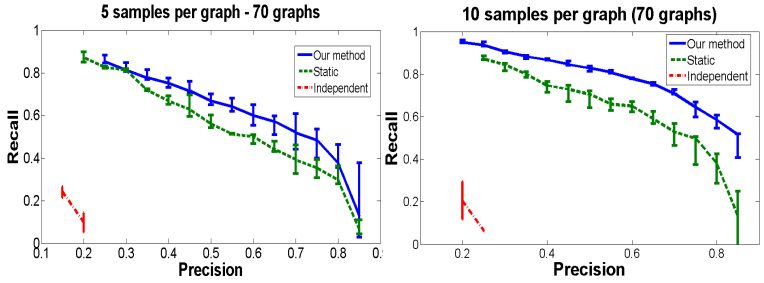

Fig. 2. Results on simulations. Our method (in blue) performs favorably to existing methods. See text for details.

4.1 Experimental Setup

We have 15 microarray measurements of 22,000 genes detailed below, that we grouped into 5 categories of 3 samples each (based on their similarities): 3 samples of S1 cells, 3 samples of T4 cells, 3 samples of T4R cells reverted by MMP inhibitors (later referred to as MMP-T4R), 3 samples of T4R cells reverted by either PI3K or MAPKK inhibitors (PI3K-MAPKK-T4R), and 3 samples of T4R cells reverted by either EGFR or integrin $\beta 1$ inhibitors (EGFR-ITGB1-T4R).

Our experimental procedure started with feature selection to reduce noise. Since some probes on Affymetrix arrays have multiple replicates, we combined measurements from these probes by taking the median, which resulted in 12,977 unique genes. Next, for each gene we calculated its median fold ratios of expression levels among each pair of the 5 groups of cells. If any of the fold ratios for a gene was greater than 1.3, it was selected for the next step. We picked 5,440 genes using this criterion.

Then, we applied *Treegl* to the 5,440 genes. In addition to taking advantage of the similarity between the T4 and T4R networks, we also explicitly penalize the difference between the S1 and T4R networks, since the T4R networks are expected to lie somewhere in between the S1 and T4. As a result, we add extra TV penalty terms between the T4R and S1 to enforce this intuition (dotted lines in Figure 3). These extra penalty terms are assigned the same parameter λ_2 as the other total variation penalties. The new optimization problem is given below ($n = 1$ corresponds to S1, $n = 2$ corresponds to T4, and $n = 3, 4, 5$ correspond to the T4R).

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(1)}_{\setminus u}, ..., \hat{\boldsymbol{\theta}}^{(n)}_{\setminus u} = \underset{\boldsymbol{\theta}^{(1)}_{\setminus u}, ..., \boldsymbol{\theta}^{(n)}_{\setminus u}}{\arg\min} \Bigg(& \sum_{n=1}^{N} \sum_{s=1}^{S_n} \left(x_u^{(n,s)} - \boldsymbol{\theta}^{(n)}_{\setminus u} \mathbf{x}^{(n,s)}_{\setminus u}\right)^2 \\ & + \lambda_1 \sum_{n=1}^{N} \|\boldsymbol{\theta}^{(n)}_{\setminus u}\|_1 + \lambda_2 \sum_{n=2}^{N} \|\boldsymbol{\theta}^{(n)}_{\setminus u} - \boldsymbol{\theta}^{(\pi(n))}_{\setminus u}\|_1 \\ & + \lambda_2 \sum_{n=3}^{N} \|\boldsymbol{\theta}^{(n)}_{\setminus u} - \boldsymbol{\theta}^{(1)}_{\setminus u}\|_1 \Bigg). \end{aligned} \quad (10)$$

All results described here are with the parameter settings of $\lambda_1 = 4$ and $\lambda_2 = 2$.

Finally, functional analysis was performed to examine genes in the identified networks. We focused our analysis on the genes in the networks which are distinct in each of the 5 groups of cell types and have positive edges. To investigate how genes involved in different biological processes interact with each other in the recovered networks, we first classified the genes in the networks into the second level Gene Ontology (GO) groups, then we used TVNViewer (http://cogito-b.ml.cmu.edu/tvnviewer/) to visualize interactions between these functional groups. Moreover, the GOstat program (Beissbarth and Speed, 2004) was employed to identify significantly enriched functional groups in the identified networks. Fisher's exact test was used by GOstat to find overrepresented functional groups among a given list of genes. Our gene universe consisted of all 12,977 genes on the arrays. A functional group was considered significant if its $p < 0.10$ with the FDR controlling procedure of Benjamini & Hochberg (Benjamini and Hochberg, 1995). We also used the GOstat program to find GO groups enriched in the subnetworks of T4 cells. A functional group was selected if its $p < 0.10$.

5 ANALYSIS OF RESULTS

5.1 Results overview

Figure 4 gives an overview of all the recovered networks using Cytoscape (Shannon *et al.*, 2003). As one can see the networks exhibit many different topologies reflecting their underlying biological differences. To shed more light on these differences, Figure 5 shows the interactions among the second level GO groups in the recovered networks. The thickness of a link between two groups is proportional to the number of edges present between genes that are members of these GO groups. T4 cells display increased activities in cell proliferation and signaling, both indicative of their malignant state, compared to S1 cells. The T4R cells lie somewhere in between: MMP-T4R cells tend to have only a few interactions, since the network is quite sparse. While both the PI3K-MAPKK-T4R and EGFR-ITGB1 networks show reduced activities in growth and locomotion compared to S1 cells, the former network has more activities in cell proliferation and reduced signaling than the latter one. Taken together, these data suggest that although T4 cells can be morphologically reverted back to the normal-looking T4R cells, the underlying molecular mechanisms in the reverted cells are different from those in either S1 or T4 cells.

5.2 GO analysis of networks

Next, we performed GO analysis to discover significantly enriched functional groups specific to each network. Our results are illustrated in Table 1.

Our data shows that highly enriched GO groups in S1 cells correspond to metabolic processes or other housekeeping functions, such as cellular respiration and DNA replication, reflecting the normal nature of these cells. On the other hand, T4 cells are enriched with genes involved in cell proliferation, growth factor activity,

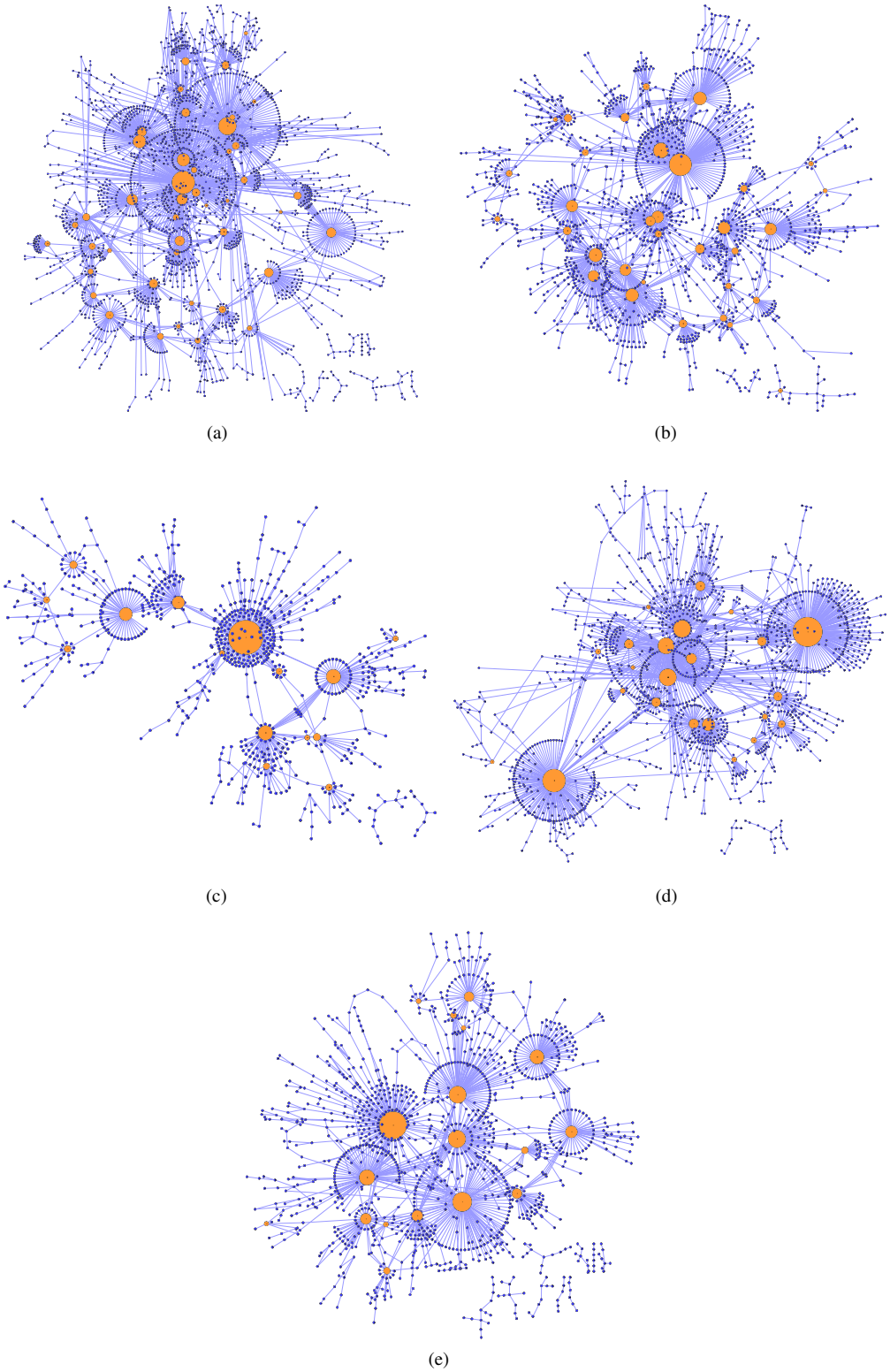

Fig. 4. Overview of the identified networks for (a) S1, (b) T4, (c) MMP-T4R, (d) PI3K-MAPKK-T4R, and (e) EGFR-ITGB1-T4R. Only edges of absolute weight > 0.1 are shown. Hubs (i.e., nodes with > 5 edges) are in orange and enlarged proportional to their degrees.

intracellular signaling cascade, angiogenesis, and actin binding group, all of which are known to play important roles in T4 as well as other cancer cells (Weaver *et al.*, 1997; Wang *et al.*, 2002; Liu *et al.*, 2004; Hanahan *et al.*, 2000). These results show that our algorithm is able to reveal what has already been known about S1 and T4 cells, and thus demonstrate the validity of our method.

Since little is known about T4R cells, we next examined the networks of the different T4R cells to gain more insight into these reverted cells. Our results show that the MMP-T4R network, like S1 cells, contains many enriched GO groups involved in metabolic processes, such as fatty acid and cofactor metabolic processes. On the other hand, however, the PI3K-MAPKK-T4R network contains genes involved mainly in post-translational protein modification, chromatin modification, thiolester hydrolase activity, and vacuole, while the EGFR-ITGB1-T4R network is predominantly overrepresented with genes participating in chromatin modification, cytoskeletal protein binding, intracellular junctions, among others. These data therefore suggest that at the molecular level T4R cells are indeed different from S1 and T4 cells, as well as from one another.

5.3 Analysis of Hubs in the T4 network

Finally, to identify potential novel drug targets in T4 cells, we examined several hubs which have high degrees as well as their neighborhood genes in these cells. Figure 6 shows the subnetworks of 5 hubs: ANXA3, CA9, HSF2BP, PTGS2, and SCG5. As expected, many of the functional gene groups enriched in the subnetworks reflect our intuition that these hubs interact closely with genes influential in cancer.

1. **ANXA3** (degree: 61) - encodes a protein belonging to the annexin family, and is known to play a role in the regulation of cell growth and is thought to be a biomarker of cancer (Jung *et al.*, 2010). In the ANXA3-subnetwork, it interacts with a number of genes related to cell proliferation, growth factor activity, and the MAP kinase signaling pathway, the latter of which is known to be one of the key signaling pathways in T4 cells (Liu *et al.*, 2004).
2. **CA9** (degree: 37) - encodes carbonic anhydrase IX. It has been implicated in cell proliferation, and has been found to be important in renal cell carcinoma (Jubb *et al.*, 2004). We see that CA9's neighborhood consists of genes involved in cell proliferation, the MAP kinase signaling pathway, golgi apparatus part, and transcription factor activity.
3. **HSF2BP** (degree: 80) - encodes heat shock transcription factor binding protein. Like the previous two hubs, HSF2BP has neighbors related to cell proliferation and the MAP kinase signaling pathway. It also has neighbors related to "response to wounding" which is known to be linked with tumorigenesis and tumor development (Fukumura *et al.*, 1998; Chang *et al.*, 2005).
4. **PTGS2** (degree: 88) - encodes prostaglandin-endoperoxide synthase 2, which is a key enzyme in prostaglandin biosynthesis. Previous evidence suggests that it is associated with risk of breast cancer (Langsenlehner *et al.*, 2006). Again, we see neighbors participating in similar activities to the previous hubs, such as cell proliferation and wound healing. Another interesting group is cell motility which suggests that the subnetwork of PTGS2 potentially plays a role in tumor cell spread. (Yamazaki *et al.*, 2005).
5. **SCG5** (degree: 78) - encodes secretogranin V, which has been found to be involved in medullary carcinoma (Marcinkiewicz *et al.*, 1988) as well as human lung cancer (Roebroek *et al.*, 1989). Again many of its neighbors are involved in cell proliferation, response to wound healing, and cell motility. Another interesting group of neighbors is those related to GTPase activity; as ras oncogenes happen to be members of the family of GTPases (Sahai and Marshall, 2002), this group of genes may also have activities implicated in cancer.

In summary, these results suggest that hubs with high degrees in the T4 network contribute to the growth, proliferation, and malignancy of T4 cells, and thus may serve as potential novel targets for breast cancer treatment.

6 DISCUSSION AND CONCLUSION

Statistically and algorithmically, the problem of estimating tree-evolving networks from multiple biological systems in the genealogy simultaneously, as solved by *Treegl*, is fundamentally different from estimating multiple networks separately from every cell type, or estimating a single "average" network from samples pooled from all cell-types (or all cell stages) in the genealogy and subsequently "trace-out" active subnetworks corresponding to each cell-type from the average network (Luscombe *et al.*, 2004), which are common practices in current system biology community. The latter two approaches either directly or indirectly assume that the network in question is a static one, and samples of nodal states, such as microarray measurements of gene expressions are *i.i.d.* within or (when pooled) across cell types. In reality, such an assumption is not only biologically invalid, but is statistically unsubstantiated and hard to leverage. First, such an assumption can lead to severe underuse of the data, and makes an already serious curse-of-dimensionality problem even harder for the following reason. Typically, in many gene expression profiling experiments, especially those from biomedical studies, the size of the sample can be extremely small (e.g., often 2-3 replica per condition or specimen) compared to the number of genes (typically $10^3 \sim 10^4$ for human) due to the difficulty of procuring many samples in laboratory experiments, which makes the directly estimated network over these genes extremely unreliable. In reality, these different cell types at different positions in the genealogy should not be drastically different, and one should expect that samples from closely related types may offer additional information to the cell type in question. Thus, estimating each point in the genealogy independently using a static reverse engineering algorithm would be largely ineffective, because there is not enough data and there are too many variables. Next, due to the presence of the genealogy that related all cell-type specific networks, the samples from all types are not identically distributed. Therefore when naively pooling them together to obtain an average network, the result may suffer from high variance, since the regulatory network could change significantly from the beginning to the end of the genealogy. The *Treegl* algorithm elegantly couples all the inference problems pertained to each network in the genealogy, and achieves a globally optimal and

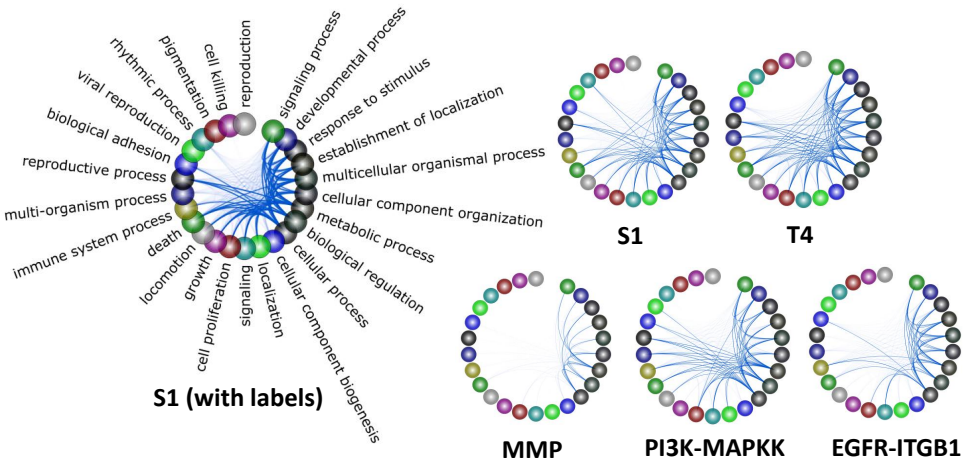

Fig. 5. Overview of results for the identified networks. Note that the nodes on the circles are not actual genes but correspond to GO process groups. The thickness of a line between two GO groups A and B is proportional to how many genes in A interact with those in B.

Table 1. Significantly enriched GO groups found in networks for (a) S1, (b) T4, (c) MMP-T4R, (d) PI3K-MAPKK-T4R, and (e) EGFR-ITGB1-T4R. Note that the T4 network contains many groups related to cell proliferation, growth, and angiogenesis, while the S1 network does not (See text for details).

(a)

GO group	p-value (FDR)
mitochondrion	2.6E-12
energy derivation by oxidation of organic compounds	1.6E-07
macromolecular metabolic process	1.8E-05
cellular respiration	1.4E-04
biopolymer metabolic process	2.7E-04
ribosome	4.0E-04
RNA metabolic process	4.1E-04
DNA replication	9.6E-02

(b)

GO group	p-value (FDR)
GTP binding	1.5E-05
cell proliferation	1.2E-03
blood vessel morphogenesis	4.8E-03
angiogenesis	1.2E-02
intracellular signaling cascade	1.5E-02
actin binding	2.1E-02
growth factor activity	9.3E-02

(c)

GO group	p-value (FDR)
mitochondrion	3.8E-10
fatty acid metabolic process	3.2E-03
cofactor metabolic process	6.0E-03
membrane enclosed lumen	1.3E-02
oxidative phosphorylation	1.3E-02
primary metabolic process	3.0E-02

(d)

GO group	p-value (FDR)
lysosomal membrane	7.4E-04
vacuole	1.2E-03
endomembrane system	8.8E-03
post-translational protein modification	4.1E-02
chromatin modification	4.1E-02
thiolester hydrolase activity	6.0E-02

(e)

GO group	p-value (FDR)
chromatine modification	2.2E-02
cytochrome-b5 reductase activity	3.3E-02
intracellular junction	4.6E-02
organelle organization and biogenesis	6.0E-02
DNA packaging	8.2E-02
cytoskeletal protein binding	8.6E-02

statistically well behaving solution based on a principled VCVS model and a convex optimization formulation.

To demonstrate our method, we applied our algorithm to a microarray dataset obtained from a progression and reversion series of breast cancer cells. Our results showed that we not only were able to identify previously known molecular signatures specific to different cell types, but also that we could provide deeper insight into the unknown molecular mechanisms underlying these cells, and therefore demonstrating the strength of our method.

Some important future directions are to consider genealogies other than a tree, and network representations beyond undirected Gaussian Graphical Models, such as a Bayesian network which is directed and can offer causal insight into the gene interactions.

ACKNOWLEDGEMENTS

We are grateful to Drs. Mina Bissell and Ren Xu for their guidance as well as providing us the cancer dataset. This research was also made possible by Grants NSF DBI-0546594, IIS-0713379, NIH 1R01GM093156, and an Alfred P. Sloan Fellowship to E.P.X.

REFERENCES

Ahmed, A. and Xing, E. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, **106**(29), 11878.

Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature genetics*, **37**(4), 382–390.

Beissbarth, T. and Speed, T. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, page 881.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bonneau, R., Reiss, D., Shannon, P., Facciotti, M., Hood, L., Baliga, N., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*, **7**(5), R36.

Bresler, G., Mossel, E., and Sly, A. (2008). Reconstruction of Markov random fields from samples: Some easy observations and algorithms. *Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques, A. Goel, K. Jansen, J. D. P. Rolim, and R. Rubinfeld, Eds., Lecture Notes in Computer Science*, **5171**(1), 343–356.

Chang, H., Nuyten, D., Sneddon, J., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y., Van't Veer, L., Bartelink, H., *et al.* (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(10), 3738.

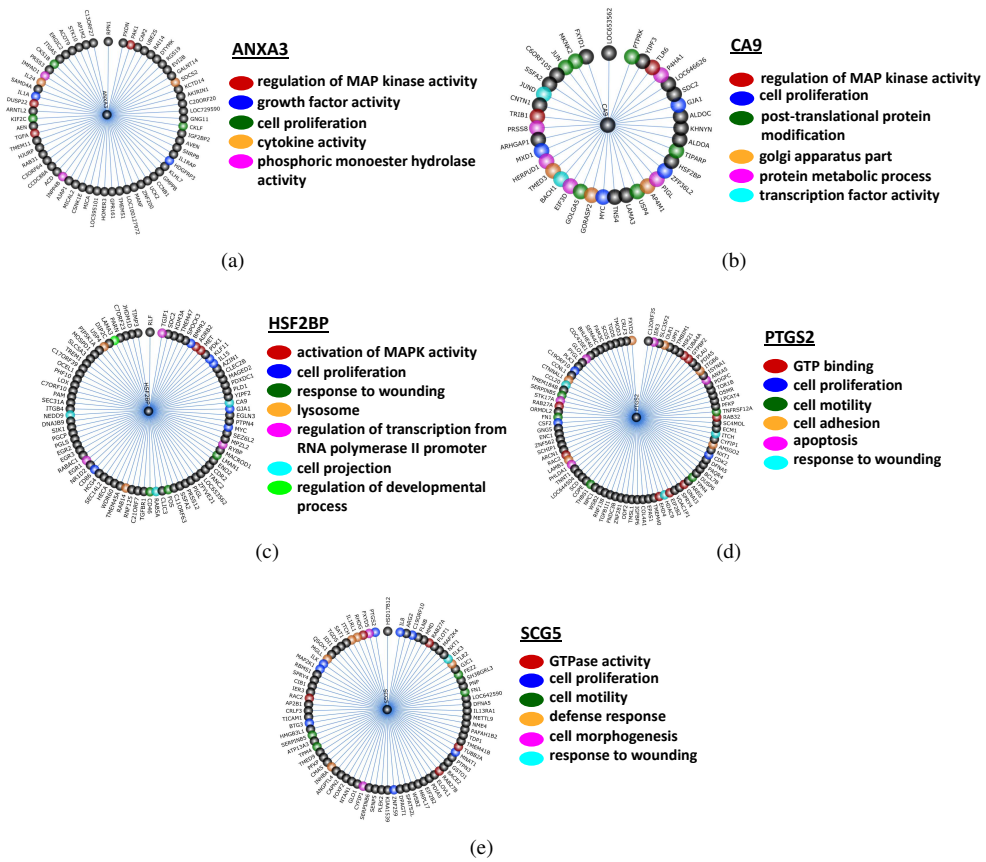

Fig. 6. Neighborhoods of a few high-degree hubs in T4 cells. A few enriched GO groups are highlighted in the subnetworks as shown.

Chen, X., Kim, S., Lin, Q., Carbonell, J., and Xing, E. (2010). Graph-Structured Multi-task Regression and an Efficient Optimization Method for General Fused Lasso. *arxiv*.

Cleveland, W. and Grosse, E. (1991). Computational methods for local regression. *Statistics and Computing*, **1**(1), 47–62.

Davidson, E. (2001). *Genomic regulatory systems*. Academic Press San Diego.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**(1), 196–212.

Fan, J. and Yao, Q. (2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. (Springer Series in Statistics). Springer.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, **7**(3-4), 601–620.

Fukumura, D., Xavier, R., Sugiura, T., Chen, Y., Park, E., Lu, N., Selig, M., Nielsen, G., Taksir, T., Jain, R., *et al.* (1998). Tumor induction of VEGF promoter activity in stromal cells. *Cell*, **94**(6), 715–725.

Grant, M., Boyd, S., and Ye, Y. (2008). CVX: Matlab software for disciplined convex programming. *Web Page and Software) [Online]. Available: http://stanford. edu/˜ boyd/cvx*.

Hanahan, D., Weinberg, R., *et al.* (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Hyatt, G., Melamed, R., Park, R., Seguritan, R., Laplace, C., Poirot, L., Zucchelli, S., Obst, R., Matos, M., Venanzi, E., Goldrath, A., Nguyen, L., Luckey, J., Yamagata, T., Herman, A., Jacobs, J., Mathis, D., and Benoist, C. (2006). Gene expression microarrays: glimpses of the immunological genome. *Nature Immunology*, **7**, 686–691.

Itoh, M., Nelson, C., Myers, C., and Bissell, M. (2007). Rap1 integrates tissue polarity, lumen formation, and tumorigenic potential in human breast epithelial cells. *Cancer Research*, **67**(10), 4759.

Jubb, A., Pham, T., Hanby, A., Frantz, G., Peale, F., Wu, T., Koeppen, H., and Hillan, K. (2004). Expression of vascular endothelial growth factor, hypoxia inducible factor 1α, and carbonic anhydrase IX in human tumours. *Journal of clinical pathology*, **57**(5), 504.

Jung, E., Moon, H., Park, S., Cho, B., Lee, S., Jeong, C., Ju, Y., Jeong, S., Lee, Y., Choi, S., *et al.* (2010). Decreased annexin A3 expression correlates with tumor progression in papillary thyroid cancer. *PROTEOMICS-Clinical Applications*, **4**(5), 528–537.

Kolar, M., Song, L., and Xing, E. (2009). Sparsistent Learning of Varying-coefficient Models with Structural Changes. *Advances in Neural Information Processing Systems*.

Langsenlehner, U., Yazdani-Biuki, B., Eder, T., Renner, W., Wascher, T., Paulweber, B., Weitzer, W., Samonigg, H., and Krippl, P. (2006). The cyclooxygenase-2 (PTGS2) 8473T>C polymorphism is associated with breast cancer risk. *Clinical Cancer Research*, **12**(4), 1392.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press, USA.

Li, Z. Chan, C. (2004). Inferring pathways and networks with a Bayesian framework. *The FASEB Journal*, **18**(6), 746–748.

Liu, H., Radisky, D., Wang, F., and Bissell, M. (2004). Polarity and proliferation are controlled by distinct signaling pathways downstream of PI3-kinase in breast epithelial tumor cells. *Journal of cell biology*, **164**(4), 603.

Lozano, A., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, **25**(12), i110.

Luscombe, N., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**(7006), 308–312.

Marcinkiewicz, M., Benjannet, S., Falgueyret, J., Seidah, N., Schurch, W., Verdy, M., Cantin, M., and Chrétien, M. (1988). Identification and localization of 7B2 protein in human, porcine, and rat thyroid gland and in human medullary carcinoma. *Endocrinology*, **123**(2), 866.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, **7**(Suppl 1), S7.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, **34**(3), 1436–1462.

Petersen, O., Rønnov-Jessen, L., Howlett, A., and Bissell, M. (1992). Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proceedings of the National Academy of Sciences*, **89**(19), 9064.

Roebroek, A., Martens, G., Duits, A., Schalken, J., van Bokhoven, A., Wagenaar, S., and Van de Ven, W. (1989). Differential expression of the gene encoding the novel pituitary polypeptide 7B2 in human lung cancer cells. *Cancer research*, **49**(15), 4154.

Sahai, E. and Marshall, C. (2002). RHO–GTPases and cancer. *Nature Reviews Cancer*, **2**(2), 133–142.

Segal, E., Wang, H., and Koller, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics-Oxford*, **19**(1), 264–272.

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498.

Song, L., Kolar, M., and Xing, E. (2009a). KELLER: estimating time-varying interactions between genes. *Bioinformatics*, **25**(12), i128.

Song, L., Kolar, M., and Xing, E. (2009b). Time-Varying Dynamic Bayesian Networks. *Advances in Neural Information Processing Systems*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Toh, K., Todd, M., and Tutuncu, R. (1999). SDPT3-A Matlab Software Package for semidefinite programming, version 2.1. *Optimization Methods and Software*, **11**, 545–581.

Wainwright, M., Ravikumar, P., and Lafferty, J. (2007). High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression. *Advances in Neural Information Processing Systems*, **19**, 1465.

Wang, F., Hansen, R., Radisky, D., Yoneda, T., Barcellos-Hoff, M., Petersen, O., Turley, E., and Bissell, M. (2002). Phenotypic reversion or death of cancer cells by altering signaling pathways in three-dimensional contexts. *Journal of the National Cancer Institute*, **94**(19), 1494.

Weaver, V., Petersen, O., Wang, F., Larabell, C., Briand, P., Damsky, C., and Bissell, M. (1997). Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. *Journal of Cell Biology*, **137**(1), 231.

Yamazaki, D., Kurisu, S., and Takenawa, T. (2005). Regulation of cancer cell motility through actin reorganization. *Cancer science*, **96**(7), 379–386.

Yeung, M., Tegnér, J., and Collins, J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the national academy of sciences of the united states of america*, **99**(9), 6163.