Leveraging Gene Networks to Estimate Perturbations on Gene Expression

Nirmalya Bandyopadhyay*, Manas Somaiya, Tamer Kahveci and Sanjay Ranka
Department of Computer Science and Engineering, University of Florida,
Gainesville, FL 32611, USA
Email: {nirmalya, mhs, tamer, ranka}@cise.ufl.edu

External factors such as radiation, drugs or chemotherapy can alter the expressions of a subset of genes. We call these genes the *primarily affected* genes. Primarily affected genes eventually can change the expressions of other genes as they activate/suppress them through interactions. Measuring the gene expressions before and after applying an external factor (i.e., perturbation) in microarray experiments can reveal how the expression of each gene changes. It however can not identify the cause of the change.

In this paper, we consider the problem of identifying primarily affected genes given the expression measurements of a set of genes before and after the application of an external perturbation. We develop a novel probabilistic method to quantify the cause of differential expression of each gene. Our method considers the possible gene interactions in regulatory and signaling networks for a large number of perturbed genes. It uses a Bayesian model to capture the dependency between the genes.

Our experiments on both real and synthetic datasets demonstrate that our method can find primarily affected genes with high accuracy. Our experiments also suggest that our method is significantly more accurate then SSEM and the Student's t-test.

1. Introduction

A significant set of microarray experiments measure gene expressions in the presence of external perturbations [7, 18]. In these perturbation experiments, radiation [36], drug [27] or other biological conditions are administered on tissues and their responses are monitored using microarrays. The expressions of the genes before perturbations correspond to *control data*, while the expressions of genes after perturbations correspond to *non-control data* [15].

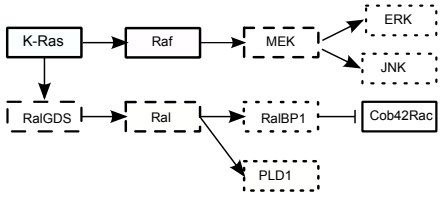


Fig. 1. Illustration of the impact of a hypothetical external perturbation on a small portion of the Pancreatic Cancer pathway. The pathway is taken from the KEGG database. The solid rectangles denote the genes affected directly by the perturbation, the dashed rectangles indicate genes secondarily affected through interactions. The dotted rectangles denote the genes that are not affected by the perturbation. $\rightarrow$ implies activation and $\dashv$ implies inhibition. In this figure, gene K-Ras, Raf and Cob42Roc are primarily affected and MEK, Ral and RalGDS are secondarily affected through interactions.

A fraction of genes respond to the external perturbation by changing their expression values significantly between control and non-control groups. Such genes are called *differentially expressed (DE)* genes [3]. The remaining genes, without noticeable changes in their expressions, are called *equally expressed (EE)* genes.

The DE genes that are directly affected by the external perturbation [11] are denoted as *primarily affected genes*. Rest of the DE genes change their expressions due to interactions with primarily affected genes and each other through signaling and regulatory networks [8, 28, 32]. We call them as *secondarily affected genes*. In this paper, the term *gene networks* is used to refer signaling and regulatory networks. Figure 1 shows the state of the genes in the Pancreatic Cancer pathway after a hypothetical external perturbation is applied [2, 34]. In this figure, genes K-Ras, Raf and Cob42Roc are primarily affected and MEK, Ral and RalGDS are secondarily affected through interactions.

We consider the problem of identifying the primarily affected genes in a perturbation experiment where gene expressions are provided before and after the application of perturbation for a set of samples. Existing methods to identify the primarily affected genes such as association analysis techniques [17, 27], haplo-insufficiency profiling [14, 13, 26] and chemical-genetic interaction mapping [30] require additional information such as fitness based assays of drug response or a library of genetic mutants. Bernardo et al. suggested a regression based approach called MNI based on the assumption that the internal

*Corresponding author.

genetic interactions are offset by the external perturbation [11]. It estimates gene-gene interaction coefficients from the control data. It then uses those coefficients to predict the target genes in the non-control data. Cosgrove et al. proposed a method named SSEM that is similar to MNI [8]. SSEM models the effect of perturbation by an explicit change of gene expression from that of the unperturbed state. Vaske et al. developed a method to infer the affected regulatory networks due to external perturbations using a graphical model called probabilistic factor graph [37]. These methods have several limitations.

(1) *Lack of gene interaction data:* The existing methods do not employ regulatory or signaling (i.e. gene networks) to model gene-gene interactions. Since gene networks are manually curated using domain experts, they are reliable sources of gene interactions. Utilizing them has the potential to more accurately solve the problem of identifying primarily affected genes.
(2) *Limited perturbations:* These methods are suitable when only a very small number of genes are perturbed, e.g., the genetic perturbation experiments are often designed for single gene perturbations [17]. However, external effects such as radiation can alter the expression of many genes directly making the existing methods to be of limited use.
(3) *Simplistic models:* Most of these methods provide only the set of genes that are directly affected by the perturbations and do not specify any error bounds. However, a non-probabilistic inference oversimplifies the problem especially in cases when a small number of gene expression measurements are available. As a result, these methods can overfit the data, making the solution unreliable.

The method we propose in this paper addresses these limitations. We assume that the underlying gene network can be modeled as a directed graph where each node represents a gene, and a directed edge from gene a to gene b represents a genetic interaction (e.g a activates or inhibits b). We define two genes as *neighbors* of each other if they share an edge. For example, in Figure 1, genes K-Ras and Raf are neighbors as K-Ras activates Raf. A neighbor can be classified as *incoming* or *outgoing* if it is at the start or at the end of the directed edge, respectively. In Figure 1, Raf is an incoming neighbor of MEK and MEK is an outgoing neighbor of Raf.

Contributions:

(1) We propose a new probabilistic method to find the primarily affected genes in microarray dataset. Our method employs gene networks as a prior belief in a Bayesian framework. When the expression level of a gene alters, it can affect some of its outgoing neighbors. Thus, the expression of a gene can change due to external perturbation or because of one or more of the affected incoming neighbors. We build our solution based on this observation. Let $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$ denote the set of all genes.
(2) We represent the external perturbation by a hypothetical gene (i.e. *metagene*) g_0 in our the gene network. An edge from the metagene the to all the other genes implies that the external perturbation has the potential to affect all the other genes. So, g_0 is an incoming neighbor to all the other genes. We call the resulting network the *extended gene network.* Our method estimates the probability that a gene g_j is DE due to an alteration in the activity of gene g_i ($g_i \in \mathcal{G} \cup \{g_0\}$, $g_j \in \mathcal{G}$) if there is an edge from g_i to g_j in the extended network. We use a Bayesian model in our solution with the help of Markov Random Field (MRF) to capture the dependency between the genes in the extended gene network.

We optimize the pseudo-likelihood of the joint posterior distribution over the random variables in the MRF using Iterative Conditional Mode (ICM) [5]. The optimization provides us the state of the genes and the pairwise causality among the genes including the metagene.

Our experiments on both real and synthetic datasets demonstrate that our method can find primarily affected genes with high accuracy. We compared our method with SSEM and Student's t test and obtained significant higher accuracy in finding out the primarily differentially expressed genes.

The rest of the paper is organized as follows. In Section 2 we describe our method in detail. In Section 3 we discuss the experiments and results. Finally, in Section 4 we describe our key conclusions.

2. Methods

In this section we describe our method. Section 2.1 presents the notations. Section 2.2 provides an overview of our solution. Section 2.3 discusses the modeling of the MRF based prior distribution. Section 2.4 describes how

we formulate a tractable approximate version of the objective function. Section 2.5 discusses how we compute the likelihood of the expression of a gene. Section 2.6 explains how we optimize the objective function.

2.1. Notation and problem formulation

We start by describing the notation used in the rest of this paper and provide a formal definition of the problem. We use two types of parameters to model this problem, namely *observed* and *hidden*. The values of observed variables are available in the given data set. We estimate the hidden variables from the observed data.

Observed variables: There are two sets of observed variables.

- **Microarray dataset:** We denote the number of microarray samples and the number of genes by N and M respectively. We represent the set of all genes in the dataset with $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$. For each gene g_i, the dataset contains the expressions before and after the perturbation (i. e. control and non-control) respectively. We denote the expressions of g_i with y_{ij} and $y_{ij}^{'}$ in control and non-control group respectively, $(1 \leq i \leq M, 1 \leq j \leq N)$. Let $\mathbf{y_i} = \{y_{i1}, y_{i2}, \cdots y_{iN}\}$ and $\mathbf{y_i}^{'} = \{y_{i1}{}', y_{i2}{}', \cdots y_{iN}{}'\}$ denote the expression values of g_i in control and non-control groups respectively. We use Y_i to denote all the data for gene g_i in control and non-control groups (i.e. $Y_i = \mathbf{y_i} \cup \mathbf{y_i}^{'}$). $\mathcal{Y} = \bigcup_{i=1}^{M} Y_i$ represents the collection of the entire gene expression data.
- **Neighborhood variables:** We use the term $\mathcal{W} = \{W_{ij}\}$ to represent if any two genes g_i and g_j are neighbors. W_{ij} $(1 \leq i, j \leq M)$ is set to 1 if g_i is an incoming neighbor of g_j (i.e. g_j has an incoming edge from g_i in the extended gene network) and 0 otherwise.

Hidden Variables: There are two sets of hidden variables.

- **State variables:** Each gene g_i can attain one of the two states (i.e. DE or EE). We introduce the variables $\mathcal{S} = \{S_i\}$ to indicate the states of the genes. Formally, S_i is DE if g_i is differentially expressed and EE if g_i is equally expressed.
- **Interaction variables:** We define the set of random variables $\mathcal{X} = \{X_{ij}\}$ to represent the joint state of genes g_i and g_j $(0 \leq i \leq M, 1 \leq j \leq M)$. Formally,

$$X_{ij} = \begin{cases} 1 \text{ if } S_i = \text{DE and } S_j = \text{DE;} \\ 2 \text{ if } S_i = \text{DE and } S_j = \text{EE;} \\ 3 \text{ if } S_i = \text{EE and } S_j = \text{DE;} \\ 4 \text{ if } S_i = \text{EE and } S_j = \text{EE;} \end{cases}$$

It is evident that the value of X_{ij} depends on the values of two independent variables S_i and S_j. Note that the values of X_{ij} are categorical in nature.

Problem formulation: We have microarray expression data $\mathcal{Y}$ and the gene network $\{\mathcal{G}, \mathcal{W}\}$ as input to the problem. From now on, the gene network $\{\mathcal{G}, \mathcal{W}\}$ will be referred to by $\mathcal{V}$. We would like to estimate the posterior density $p(X_{ij}|\mathcal{X} - X_{ij}, \mathcal{Y}, \mathcal{V}, W_{ij} = 1)$. Specifically, a lower value of $p(X_{0j} = 2|\mathcal{X} - X_{0j}, \mathcal{Y}, \mathcal{V}, W_{ij} = 1)$ indicates a higher chance that the gene g_j is primarily affected, as X_{0j} = 2 indicates that the metagene is DE and gene g_j is EE. Based on this probability estimation, we create a list of primarily affected genes.

2.2. Overview of our solution

An approach to solve our problem can be to maximize a likelihood distribution over the gene expression $\mathcal{Y}$ where $\mathcal{X}$ are the parameters of the distribution. The objective is to obtain the maximum likelihood estimate (MLE) of $\mathcal{X}$. However, there are two problems in this this approach. First, MLE requires a large number of data points to accurately estimate the parameters. Second, MLE depends only on the observed data and cannot utilize domain specific knowledge; as a result it leads to overfitting of data and poor generalization.

We develop a Bayesian framework for estimating $\mathcal{X}$ that addresses the above-mentioned limitations of the existing approaches. Bayesian approaches can generally estimate the parameters with fewer data-points, which makes our approach more suitable for perturbation experiments [6].

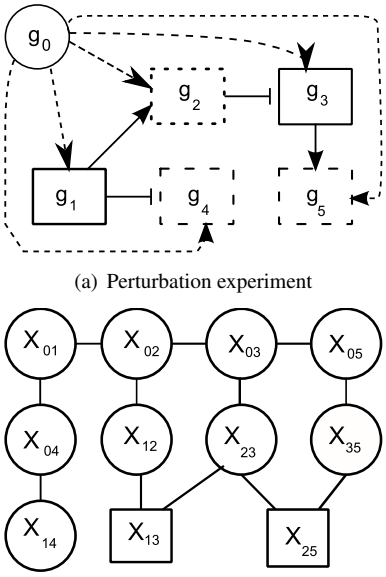
We estimate the probability of X_{ij} given the other observed and hidden variables. In this approach, we aim to maximize the joint density of the $\mathcal{X}$ variables given the gene expressions $\mathcal{Y}$ and the gene network $\mathcal{V}$. Thus, the objective to maximize is given by,

$$P(\mathcal{X}|\mathcal{Y}, \mathcal{V}, \theta_Y, \theta_X) = \frac{P(\mathcal{Y}|\mathcal{X}, \mathcal{V}, \theta_Y) P(\mathcal{X}|\mathcal{V}, \theta_X)}{\sum_{\mathcal{X}} P(\mathcal{Y}|\mathcal{X}, \mathcal{V}, \theta_Y) P(\mathcal{X}|\mathcal{V}, \theta_X)} \quad (1)$$

θ_Y is the set of parameters for the likelihood function $P(\mathcal{Y}|\mathcal{X}, \mathcal{V}, \theta_Y)$ and θ_X is the set of parameters for the prior density function $P(\mathcal{X}|\mathcal{V}, \theta_X)$. θ_X and θ_Y will be discussed in Sections 2.3 and 2.5 respectively.

Since a direct optimization of Equation 1 is impractical due to exponential number of terms in the denominator, we define a more tractable objective function as discussed in Section 2.4. We use iterative conditional mode (ICM) to optimize the objective function and obtain an assignment of $\mathcal{X}$, θ_X and θ_Y. Finally we estimate the posterior probability $p(X_{ij}|\mathcal{X} - X_{ij}, \mathcal{Y}, \theta_X, \theta_Y)$ for every X_{ij} when W_{ij} = 1. Using this posterior probability, we quantify the chance that one gene is DE due to one of its incoming neighbors.

2.3. Computation of the prior density function



(a) Perturbation experiment

(b) Markov random field graph

Fig. 2. (a) A small hypothetical gene network with a perturbation. The circle g_0 represents the abstraction of the external perturbation i.e. g_0. Rectangles denote genes. $\rightarrow$ implies activation and $\dashv$ implies inhibition. The dotted arrow from g_0 indicates potential effect on each genes. The directly impacted DE genes g_1 and g_3 are denoted by solid rectangle. Dashed rectangles g_4 and g_5 imply secondarily impacted DE genes. Dotted rectangle is for the EE gene g_2. (b) The graph for Markov random field created from the hypothetical gene network in (a). For each neighbor pair we create a circular node. We create two rectangular nodes that do not correspond to any neighbor pair, however they are part of the MRF graph. Two nodes are connected with an undirected edge if they share a subscript at same position and the two genes corresponding to the other subscript interact in the gene network. Also, at least one of the nodes represents an interaction in the network. For example, node X_{04} and X_{14} are connected as they share 4 at second position and g_0 is an incoming neighbor of g_1. Both X_{04} and X_{14} correspond to interactions in the network.

In this section, we describe how we build the prior density function $P(\mathcal{X}|\theta_X)$. We incorporate information from biological networks as prior belief in this density function. The following two assumptions encapsulate our belief about gene interactions.

(1) Each gene can affect the expressions of its outgoing neighbors. If the activity of a gene is altered, the effect can propagate to its outgoing neighbors.
(2) The metagene g_0 (i.e. external perturbation) can affect the expression of every other gene. This is easy to visualize as the external perturbation such as radiation can change the activity of any of the genes.

Clearly, when the data does not follow one or more of the hypotheses, the optimization function can overcome the prior belief with a strong support from the data.

In order to compute the prior density function, we define a Markov Random Field (MRF) over the $\mathcal{X}$ variables [24]. MRF is a probabilistic model, where the state of a variable depends only on the states of its neighbors. MRF is useful to model our problem as the states of genes depend on their neighbors. Here, the MRF is an undirected graph Ψ = ($\mathcal{X}$, $\mathcal{E}$), where $\mathcal{X}$ = $\{X_{ij}\}$ variables represent the vertices of the graph (i.e. each interaction variable X_{ij} corresponds to a vertex). We denote the set of edges with $\mathcal{E}$ = $\{(X_{ij}, X_{pj})|W_{pi} = W_{ij} = 1\} \cup \{(X_{ij}, X_{ik})|W_{jk} = W_{ij} = 1\}$. Thus, two variables in $\mathcal{X}$ share an edge if they share a common subscript at the same position, the two genes corresponding to the other subscript interact in the gene network and finally at least one of the two nodes represents an interaction in the extended gene network. For example, in Figure 2(b), X_{35} and X_{25} are neighbors, as they share 5 (i.e. gene g_5) as the second subscript and g_2 and g_3 interact in the gene network in Figure 2(a). Also, X_{35} corresponds to a real interaction in the network.

One important point to note is that, this graph does not use the state variables $\mathcal{S}$ to model the dependencies between the genes. Rather, it establishes those dependencies over the $\mathcal{X}$ variables. For example, in Figure 2(b) we draw the MRF graph corresponding to the hypothetical gene network in Figure 2(a). In the gene network, there is an edge from g_2 to g_3. So, g_2 can potentially change the state of g_3. We create an edge from X_{12} to X_{13} that corresponds to the edge from g_2 to g_3. As g_1 is common for X_{12} and X_{13}, if they assume the same value (i.e. $X_{12} = X_{13}$), it implies that the genes g_2 and g_3 are in the same state (i.e. $S_2 = S_3$). We formulate these dependency constraints using a set of unary and bi-

nary functions called *feature functions*. We discuss these feature functions next.

We denote the neighbors of X_{ij} in the MRF graph as $X^*_{ij} = \{X_{pj}|W_{pi} = 1\} \cup \{X_{ik}|W_{jk} = 1\}$. We define a clique over each X_{ij} and its neighbors X^*_{ij} by C_{ij} provided $W_{ij} = 1$. A feature function $f(C_{ij})$ is a Boolean function defined over the cliques C_{ij} of Ψ. This function evaluates to 1 or 0, if it is satisfied or not, respectively. We define a *potential function* $\psi(C_{ij})$ corresponding to $f(C_{ij})$ as an exponential function given by $\exp(\gamma f(C_{ij}))$. Here γ is a coefficient associated with $f(C_{ij})$ that represents the relevance of $f(C_{ij})$ in the MRF. According to Hammersley-Clifford theorem, we express the joint density function of the MRF over $\mathcal{X}$ as product of potential functions constructed for that MRF as, $p(\mathcal{X}|\theta_X) = \frac{1}{\Delta}\prod_{C_{ij},W_{ij}=1}\psi(C_{ij})$ [16]. In this formulation, Δ is the normalization function $\Delta = \sum_{\mathcal{X}}\prod_{C_{ij},W_{ij}=1}\psi(C_{ij})$. To limit the complexity of our model, we consider only cliques of size one and two.

We define four feature functions to capture the dependencies among the variables in $\mathcal{X}$ according to the two hypotheses. Based on the number of input variables, they are classified as *unary* and *binary* feature functions.

Table 1. The table enumerates the truth values for the two binary feature functions. Only the permitted entries are annotated with 0/1. The blank entries corresponds to combinations that are not possible. (a) $f_3(X_{ij}, X_{pj})$ represents the feature function for *left equality*. (b) $f_4(X_{ij}, X_{ik})$ represents the feature function for *right equality*.

X_{ij} \ X_{pj}	1	2	3	4
1	1		0	
2		1		0
3	0		1	
4		0		1

(a) $f_3(X_{ij}, X_{pj})$

X_{ij} \ X_{ik}	1	2	3	4
1	1	0		
2	0	1		
3			1	0
4			0	1

(b) $f_4(X_{ij}, X_{ik})$

Unary feature functions: A primary component of the prior density function is modeling the frequency of X_{ij} itself. We capture this frequency using two unary feature functions defined over singleton cliques. We define a feature function $F_1(X_{ij})$ which returns one when $X_{ij} = 1$ and 0 otherwise. To capture the complemented events, we define another feature function $F_2(X_{ij})$, which returns to 1 when $X_{ij} = 0$ and returns 0 otherwise.

Binary feature functions: These feature functions are defined to incorporate the two assumptions stated at the beginning of this section. Consider a sequence of four genes g_1, g_2, g_3 and g_5 in Figure 2(a). X_{23} is a variable in the MRF graph that depends on the states of g_2 and g_3. X_{13} is a neighbor of X_{23} in MRF graph as g_1 is an incoming neighbor of g_2 in the gene network. Similarly, X_{25} is a neighbor of X_{23} as g_5 is an outgoing neighbor of g_3. If S_1 equals to S_2 then $X_{23} = X_{13}$. Similarly if S_3 equals to S_5 then $X_{23} = X_{25}$. We capture these events in two feature functions for X_{ij} based on the incoming neighbors of g_i and the outgoing neighbors of g_j.

- **Left equality:** Let us denote the incoming neighbors of g_i with $In(g_i)$. We write a feature function $f_3(X_{ij}, X_{pj}), \forall p, g_p \in In(g_i)$. $f_3(X_{ij}, X_{pj}) = 1$ if $S_i = S_p$ and $W_{pi} = W_{ij} = 1$. Otherwise, $f_3(X_{ij}, X_{pj}) = 0$. We denote the summation of this function over all the incoming neighbors of g_i as,

$$F_3(X_{ij}) = \sum_{p,W_{ij}=1,W_{pi}=1} f_3(X_{ij}, X_{pj}).$$

- **Right equality:** Let us denote the outgoing neighbors of g_j as $Out(g_j)$. We define a feature function $f_4(X_{ij}, X_{ik}), \forall k, g_k \in Out(g_j)$. $f_4(X_{ij}, X_{ik}) = 1$ if $S_k = S_j$ and $W_{jk} = W_{ij} = 1$. Otherwise, $f_4(X_{ij}, X_{ik}) = 0$. We denote the summation of this function over all the outgoing neighbors of g_j as,

$$F_4(X_{ij}) = \sum_{k,W_{ij}=1,W_{jk}=1} f_4(X_{ij}, X_{ik}).$$

Table 1 enumerates the truth values of the binary feature functions for different values of their arguments. Only the permitted entries are annotated with zero and one. The other entries require illegal combination of argument values.

In the binary feature functions X_{pj} or X_{ik} may not represent any interactions from the extended gene network when $W_{pj} = 0$ or $W_{ik} = 0$, respectively. We represent them by rectangles in Figure 2(b).

Based on these feature functions, we define the joint density function of $\mathcal{X}$ as,

$$p(\mathcal{X}|\theta_X) = \frac{1}{\Delta} exp(\sum_{i,j,W_{ij}=1,k\in\{1,2,\cdots,4\}} \gamma_k F_k(X_{ij})) \quad (2)$$

In the above equation γ_k, $k \in \{1, 2, \cdots 4\}$ are the coefficients of the four feature functions in MRF. In the next section, we discuss how we define the objective function with respect to the MRF. We also describe how we formulate the posterior probability density function for X_{ij}.

2.4. Objective function approximation

A direct maximization of the objective function given by Equation 1 is impractical, as it requires evaluation of exponential number of terms in the denominator. We employ pseudo-likelihood as an established substitute to Equation 1 [4]. Pseudo-likelihood is the simple product of the conditional probability density function of the X_{ij} variables. Geman et al. proved the consistency of the maximum pseudo-likelihood estimate [12]. The approximated objective function can be written as,

$$F = \arg\max_{\mathcal{X}}(\prod_{i,j} F_{ij}) \tag{3}$$

We derive the posterior density function F_{ij} of X_{ij} when W_{ij} = 1 as,

$$\begin{aligned} & F_{ij} \\ & = p(X_{ij}|\mathcal{X} - X_{ij}, \mathcal{Y}, \theta_X, \theta_Y, W_{ij} = 1) \\ & = \frac{p(Y_i, Y_j|X_{ij}, X^*_{ij}, \theta_Y, W_{ij} = 1)p(X_{ij}|\mathcal{X} - X_{ij}, \mathcal{Y}, \theta_X, W_{ij} = 1)}{\sum_{X_{ij} \in \{1,\cdots,4\}} p(Y_i, Y_j|X_{ij}, X^*_{ij}, \theta_Y, W_{ij} = 1)} \end{aligned} \tag{4}$$

There are two different terms in objective function of Equation 4. $p(X_{ij}|\mathcal{X} - X_{ij}, \mathcal{Y}, \theta_X, \theta_Y, W_{ij} = 1)$ stands for the conditional prior density function of X_{ij} which can be derived from Equation 2 using Bayes rule. We discuss $p(Y_i, Y_j|X_{ij}, X^*_{ij}, \theta_Y, W_{ij} = 1)$, the likelihood function in the next section.

2.5. Calculation of likelihood density function

In this section, we describe how we derive the likelihood function. We assume that gene expressions in a group follow a normal distribution. We can redo the derivations if gene expressions follow some other distribution.

Consider a set of measurements for a gene g_i that follows a single Gaussian distribution by $\mathbf{z_i} = \{z_{i1}, z_{i2}, \cdots, z_{iN}\}$. We denote the latent mean of $\mathbf{z_i}$ by μ and the standard deviation by σ. As different genes can have different average expressions, we assume that μ follows a genome-wise normal distribution with mean μ_0 and standard deviation τ [23]. Thus, for $\mathbf{z_i}$, the likelihood for the data points in that group is given by,

$$\begin{aligned} L(\mathbf{z}|\mu_0, \sigma^2, \tau^2) &= \int [\prod_{i=1}^{n} \mathcal{N}(z_i|\mu, \sigma^2)]\mathcal{N}(\mu|\mu_0, \tau^2)d\mu \\ &= \frac{\sigma}{(\sqrt{2\pi}\sigma)^n\sqrt{n\tau^2 + \sigma^2}} exp(-\frac{\sum_i z_i^2}{2\sigma^2} - \frac{\mu_0^2}{2\tau^2}) \cdot \\ & \quad exp(\frac{\frac{\tau^2 n^2 \bar{z}^2}{\sigma^2} + \frac{\sigma^2\mu_0^2}{\tau^2} + 2n\bar{z}\mu_0}{2(n\tau^2 + \sigma^2)}) \end{aligned} \tag{5}$$

The reader can find the derivation of Equation 5 in Demichelis et al [10].

If a gene is DE, its expression measurements in control and non-control groups follow separate distributions. On the other hand, for equally expressed genes, all the measurements in both the groups share the same mean. The data likelihood for a DE gene is given by,

$$\mathcal{L}(g_i) = \begin{cases} L(\mathbf{y_i}|\mu_0, \sigma^2, \tau^2)L(\mathbf{y_i'}|\mu_0, \sigma^2, \tau^2), \text{if } S_i = DE. \\ L(\mathbf{y_i} \cup \mathbf{y_i'}|\mu_0, \sigma^2, \tau^2), \text{if } S_i = EE \end{cases} \tag{6}$$

Now we are ready to derive the likelihood density for different values of X_{ij}. Let us denote the set of parameters $\{\mu, \sigma, \tau\}$ by θ_Y.

We have four different forms for the likelihood of (Y_i, Y_j) due to four different values it can assume. However, we shall derive only for X_{ij} = 1, as for the other values of X_{ij} we have similar derivations.

$$\begin{aligned} & p(Y_i, Y_j|X_{ij} = 1, X^*_{ij}, \theta_Y, W_{ij} = 1) \\ & = \sum_{\tau_i, \tau_j \in \{DE, EE\}} p(Y_i, Y_j|S_i = \tau_i, S_j = \tau_j, \theta_Y, W_{ij} = 1) \cdot \\ & \quad p(S_i = \tau_i, S_j = \tau_j|X_{ij} = 1, X^*_{ij}, \theta_Y, W_{ij} = 1) \end{aligned} \tag{7}$$

From the definition of X_{ij}, $p(S_i = \tau_i, S_j = \tau_j|X_{ij} = 1, X^*_{ij}, \theta_Y)$ equals to 1 when S_i = DE and S_j = DE. Its value is zero for all other values of S_i and S_j. So, continuing from the last step of Equation 7,

$$\begin{aligned} & p(Y_i, Y_j|X_{ij} = 1, X^*_{ij}, \theta_Y, W_{ij} = 1) \\ & = p(Y_i, Y_j|S_i = DE, S_j = DE, \theta_Y, W_{ij} = 1) \\ & = p(Y_i|S_i = DE, S_j = DE, \theta_Y) \cdot \\ & \quad p(Y_j|S_i = DE, S_j = DE, \theta_Y) \\ & = p(Y_i|S_i = DE, \theta_Y)p(Y_j|S_j = DE, \theta_Y) \\ & = \mathcal{L}(g_i)\mathcal{L}(g_j) \end{aligned}$$

In a similar way, we can derive the likelihood functions for the other three values of X_{ij} variable. A special case arises when g_i is the metagene, i.e. g_0. Specifically, $\mathcal{L}(g_0)$ = 1 if S_0 = DE and 0 otherwise, as, according to our assumption the metagene is always DE.

2.6. Objective function optimization

So far, we have described how we compute the posterior density function. The final challenge is to find the values of the hidden variables that maximize the objective function (Equation 3). We develop an iterative algorithm to address this challenge.

In our model, we have three different sets of parameters. The nodes of the MRF given by $\mathcal{X}$ consist of one set. Other two sets are the parameters of conditional probability density function of X_{ij} and likelihood function of observed data given by $\theta_X = \{\gamma_1, \cdots \gamma_4\}$ and $\theta_Y = \{\mu_0, \sigma, \tau\}$, respectively. In each iteration, we first estimate θ_X and θ_Y based on the estimated value of $\mathcal{X}$ in the previous iteration. Next, based on the estimated parameters, we estimate $\mathcal{X}$ that maximize the objective function in Equation 3.

The likelihood function is non-convex in terms of the parameters $\theta_Y = \{\mu_0, \sigma, \tau\}$. Also, the conditional density is non-convex in terms of $\theta_X = \{\gamma_1, \cdots \gamma_4\}$. We use a global optimization method called differential evolution to optimize both of them [35]. To optimize the objective function in equation 3, we employ the ICM algorithm described by Besag [5]. Briefly, our iterative algorithm works as follows.

(1) Obtain an initial estimate of $\mathcal{S}$ variables. In our implementation we use student's t-test assuming the data follows normal distribution. We use 5% confidence interval for this purpose.

(2) Estimate parameters θ_Y that maximizes the data likelihood function given by,

$$\arg\max_{\theta_Y} \prod_{X_{ij}} p(Y_i, Y_j | X_{ij}, X^*_{ij}, \theta_Y, W_{ij} = 1)$$

We implement this step using Differential Evolution, which is similar to the genetic algorithm.

(3) Calculate an estimate of the parameters θ_X that maximizes the conditional prior density function by,

$$\arg\max_{\theta_X} \prod_{X_{ij}} p(X_{ij} | \mathcal{X} - \{X_{ij}\}, \theta_X, W_{ij} = 1)$$

We also implement this step using Differential Evolution.

(4) Carry out a single cycle of ICM using the current estimate of $\mathcal{S}$, θ_X and θ_Y. For all S_i, maximize $\prod_{X_{mn}} p(X_{mn} | \mathcal{X} - X_{mn}, \mathcal{Y}, \theta_X, \theta_Y, W_{mn} = 1)$ when $X_{mn} \in \{X_{rt} | r = i \text{ or } t = i\}$ and $W_{rt} = 1$.

(5) Go to step 2 for a fixed number of cycles or until $\mathcal{X}$ converges to a certain predefined value.

We optimize the objective function in terms of the S_i ($1 \leq i \leq M$) variables instead of X_{ij} variables. Specifically, in step 4, we go over all the S_i variables, and optimize F_{ij} function (given by Equation 4) for only those X_{ij} variables that are impacted by the change of S_i. The optimization procedure is guaranteed to converge since in every iteration the value of the objective function increases. We continue the iterative process, until the changes in estimates of the parameters between two consecutive iterations reach below a certain cutoff level.

3. Experiments

In this section we discuss the experiments we conducted to evaluate the quality of our method. We implemented our method in MATLAB and Java. We obtained an implementation of Differential Evolution from the `http://www.icsi.berkeley.edu/~storn/code.html`. We compared our method with SSEM [8] as SSEM is one of the most recent methods that can be used to solve the problem considered in this paper. We obtained SSEM from `http://gardnerlab.bu.edu/SSEMLasso`. We ran our code on an AMD Opteron 2.4 Ghz workstation with 4GB memory.

Dataset: We used the dataset collected by Smirnov et al. [33]. It was generated using 10 Gy ionizing radiation over immortalized B cells obtained from 155 members of 15 Centre d'tude du Polymorphisme Humain (CEPH) Utah pedigrees [9]. Microarray snapshots were obtained at 0th hour (i.e., before the radiation) and 2 and 6 hours after the radiation. We adapt the time series data to create the control and non-control data for our experiments. We use the data before radiation as control data. For the non-control data we calculate the expected expressions of a gene at each points after the radiation. We select the one with higher absolute difference from the expected expression of control data for that gene. This dataset is used in the experiments described in Sections 3.1 and 3.2. For the experiments described in Sections 3.3 and 3.4, we derive new datasets using this data. The details of this process can be found in corresponding sections.

We also collect 24,663 genetic interactions from the 105 regulatory and signaling pathways of KEGG database [22]. Overall 2,335 genes belong to at least one pathway in KEGG. We consider only the genes that take part in the gene networks in our model.

Table 2. List of top 25 genes that are mostly affected by external perturbation. The dataset was generated using 10 Gy ionizing radiation over immortalized B cells obtained from 155 members of 15 Centre d'tude du Polymorphisme Humain (CEPH) Utah pedigrees Genes are tabulated row-wise, in increasing order of ranking.

PGF	IL8RB	FOSL1	F2R	PPM1D
MDM2	CDKN1A	TNC	PLXNB2	EPHA2
DDB2	TP53I3	PLK1	TNFSF9	ADRB2
MAP3K12	JUN	SORBS1	LRDD	SDC1
MYC	PRKAB1	EI24	DDIT4	FAS

3.1. Biological significance

In this section, we investigate the support in existing literature for susceptibility to radiation based perturbation for the primarily affected genes found by our method. We train our method on the dataset described above. After the optimization we rank each gene g_j in decreasing order of $\mathcal{L}(g_0)\mathcal{L}(g_j)$, where $\mathcal{L}(g_j)$ is given by Equation 6. We tabulate the top 25 genes in Table 2.

Nine out of the ten highest ranked genes have significant biological evidence that they are impacted by radiation. Imaoka et al. [19] compared the gene expression between normal mammary glands to spontaneous and γ-radiation induced cancerous glands of rat. The PGF (parental growth factor) gene showed differential expression in both spontaneous and irradiated carcinomas. Nagtegaal et al. [29] applied radiation to human rectal adenocarcinoma and compared the gene response to that of normal tissues. The cytokines and receptor IL8RB showed differential expression between normal and irradiated rectal tissues. Amundson et al. [1] administered γ-radiation to p-53 wild type ML-1 human myeloid cell line. FOSL1 (known by FRA1 that time) showed differential expression as the stress response. Lin et al. [25] applied ionizing radiation on human lymphoblastoid cells. F2R, a coagulation factor II receptor, was upregulated in that experiment. Jen et al. [21] investigated the effect of ionizing radiation on the transcriptional response of lymphoblastoid cells in time series microarray experiments. PPM1D, a gene related to DNA repair, showed response to both 3Gy and 10Gy radiation. Wu et al. [38] conducted a high dose UV radiation experiment to observe the relation between MDM2 gene on p53 gene. Their experiment revealed that initially both protein and mRNA level of MDM2 increases in a p53 independent manner, which clearly substantiated the direct effect of radiation on MDM2. Jakob et al. [20] irradiated human fibroblasts with accelerated lead ions. Confocal microscopy discovered a single, bright focus of CDKN1A protein in the nuclei of human fibroblast within 2 minutes after radiation. Rieger et al. [31] applied both ultra violet and infrared radiation on fifteen human cell lines and observed that PLXNB2 was up-regulated for both kind of radiations. Zhang et al. [39] reported that EPHA2 worked as an essential mediator of UV-radiation induced apoptosis.

This experiment demonstrates that we find sufficient support in existing literature that the top ranked genes found by our method (i.e. highly likely to be primarily affected) are affected by radiation.

3.2. Evaluation of the rankings of neighbor genes

Recall that our goal is to find the primarily affected genes. We achieve this objective by computing the probability for each DE gene to contribute towards the change in the expression of its outgoing neighbors. In this experiment, we evaluate our success in terms of how accurately we rank the contribution probabilities of the genes as discussed in the next paragraph.

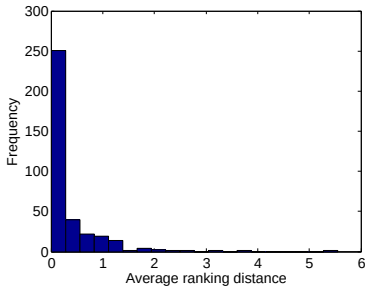


Fig. 3. Frequency of average distance of rankings over training and testing data. The figure shows that the difference is very close to zero. This suggests that our method can rank the probabilistic effect of the incoming neighbors of the genes with great precision. The average difference between the ranks obtained in the training and the testing data is less than one position in 92.7% of the cases.

We divide the dataset of 155 samples into training and testing set in 2:1 ratio. We create a ranked list for each DE gene as follows. For each DE gene, we sort its incoming DE neighbors in decreasing order of their data likelihood probability with respect to the outgoing neighbor. For example, assume g_1 to be DE. It has four incoming DE neighbors g_0, g_2, g_3 and g_4 where g_0 is the metagene. Let NL_{ij} denotes the normalized likelihood function $\frac{p(Y_i,Y_j|X_{ij}=1,X^*_{ij},\theta_Y)}{\sum_{X_{ij}\in\{1,2,3,4\}} p(Y_i,Y_j|X_{ij},X^*_{ij},\theta_Y)}$ of X_{ij}. For instance, If $NL_{01} \geq NL_{41} \geq NL_{21} \geq NL_{31}$, then the sorted list is $\{g_0, g_4, g_2, g_3\}$. We denote the sorted list

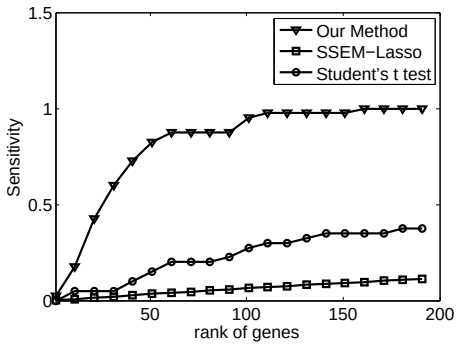
as a ranking of the incoming DE neighbors. Let us denote the position of a gene g_i in the ranking of g_j for training data $\rho_{g_j}(g_i)$. We create another set of rankings from the testing data likelihood probability. Let us denote the position of g_i in the ranking of g_j from testing data by $\rho^{'}_{g_j}(g_i)$. For a gene g_j we define the *average ranking distance* between training and testing data as $\delta(g_j) = \frac{\sum_{g_i \in IN(g_j)} abs(\rho_{g_j}(g_i) - \rho^{'}_{g_j}(g_i))}{|IN(g_j)|}$, where $IN(g_j)$ is the set of incoming DE neighbors for g_i, abs(.) denotes the absolute value and $|IN(g_j)|$ stands for the cardinality of $IN(g_j)$.

We calculated the average ranking distance for all the genes that have incoming neighbors apart from the metagene. This experiment was repeated three times with a different set of training and testing data. We create a histogram for the average differences from the three experiments in Figure 3. It shows that the difference in average ranking distance is very close to zero. The average difference between the ranks obtained in the training and the testing data is less than one position in 92.7% of the cases. Thus, we have demonstrated that we can accurately rank the contribution probabilities of incoming neighbors for DE genes in test dataset based on the model parameters learned from the training dataset.

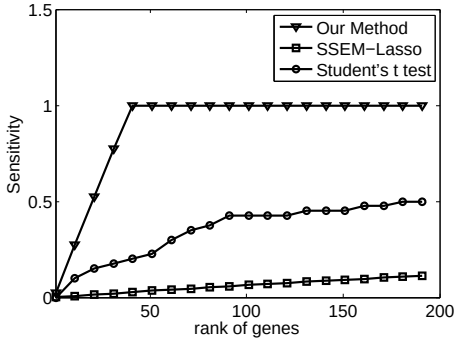
3.3. Comparison to other methods

In this section, we compare the accuracy of our method to that of SSEM and a simpler method *Student's t test*.

Synthetic data generation: We simulated real perturbation events to prepare synthetic data with known primarily and secondarily affected genes in a controlled setting. We use the gene network derived from KEGG first to select a random gene from the network and denote it as a primarily affected DE gene. We traverse the ancestors in a breadth first manner. For each of the ancestor, we made it a secondarily affected DE gene with a probability of $1-(1-q)^{\eta}$, where η is the number of incoming DE neighbors. Here q (0.4 in our experiments) is the probability that a gene is DE due to a DE predecessor. We repeat these steps to create the desired number of primarily affected genes. After the classification of the genes we create control and non-control data for each of them for over N patients. We use the control part of the real dataset in Smirnov et al. [33] as the control part of our synthetic dataset. To generate the non-control dataset, we traverse each of the genes that participate in the gene networks. Suppose, for a gene g_i, the mean and standard deviation of its expression in the control dataset are given by μ_i and σ_i respectively. If the gene is EE we generate its non-control data points from the a normal distribution given by the parameters (μ_i, σ_i^2). If the gene is DE, we use the same variance as that of the control group. However, we use a different mean. For the primarily and secondarily affected genes we use $\mu_i^{'} = \mu_i \pm d_p$ and $\mu_i^{'} = \mu_i \pm d_s$ respectively, where $d_p > d_s$.



(a) Gap = 0.2 $\times\sigma$



(b) Gap = 0.6 $\times$ σ

Fig. 4. Comparison of our method to SSEM and t-test. The number of primarily affected genes is 50. The gap between the mean of primarily affected and secondarily affected genes are 0.2 to 0.6 $\times\sigma$, where σ is estimated from the real dataset. The figures indicate that our method outperforms SSEM and t-test.

Experimental setup: Given an input dataset, using each of the three methods, we ranked all the genes. Highly ranked genes have higher chance of being a primarily affected gene according to each method. We explain how we do the ranking in the following.

- **Our method:** We sort the genes in decreasing order of likelihood with the metagene. A higher likelihood value implies a higher chance of being primarily affected.
- **SSEM:** We train SSEM on the control dataset, where

it learns the correlation between the genes. We test SSEM on the non-control dataset, where it produces a rank for each single data point.

- **Student's t test:** We used the function called $ttest2$ from MATLAB. We apply it on every individual gene, where it takes control and non-control dataset as input and produces a p-value as output. By default, null hypothesis is that "the differences of two input data set are a random sample from a normal distribution with mean 0 and unknown variance, against the alternative that the mean is not 0". Thus, the null hypothesis corresponds to the assumption that the gene is EE. So a substantially lower p-value implies a higher chance of being primarily affected. We performed the test on all the genes and rank them according the increasing order of p-values.

Let us assume the set of primarily affected genes as PG and first k elements of the ranking as RG_k. We define the sensitivity of the ranking at position k by $\eta_k = \frac{|PG \cap RG_k|}{|PG|}$. Thus, a higher value of η_k denotes a higher sensitivity. We prepare a sensitivity vector $\{\eta_1, \eta_2, \cdots \eta_{|R|}\}$, by arraying the sensitivity of a ranking at all the positions of the ranks. Here, $|R|$ denotes the cardinality of the ranking. For SSEM we obtain a sensitivity vector for every data points in the non-control dataset. We create a consolidated sensitivity vector by averaging them.

Results: We conducted experiments by for $\frac{d_s - d_p}{\sigma} = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5, 1.75\}$, number of primarily affected genes = $\{10, 50\}$ and number of data points = $\{10, 20, 40, 60, 80, 100, 125, 155\}$. Here, σ corresponds to the standard deviation of the expressions of genes in the dataset. However, due to space limitation we discuss only two of them in this paper (see Figure 4). The results we discuss correspond to the cases when we have 50 primarily affected genes and 155 data points. The results of the other experiments are similar to those in Figure 4(b).

Figures 4(a) and 4(b) show the sensitivity of the three methods when $(d_s - d_p) = 0.2 \times\sigma$ and $0.6 \times\sigma$ respectively. The former one corresponds to the computationally harder case as the difference between the control and non-control datasets is small. As the gap between d_s and d_p increases identifying primarily affected genes becomes easier.

From the figure, we observe that our method is significantly more sensitive than the other two methods for all datasets consistently. It reaches high sensitivity (more than 90%) using the top 150 ranked genes when the gap is small, and using the top 50 genes as the gap increases to $0.6 \times\sigma$. The results were similar for larger gap values (results not shown). The t test reaches around 40% and 50% sensitivity at 200 ranking position respectively. SSEM's sensitivity is below 0.25 for all experiments even within the top 200 positions.

We believe that there are two major factors for improved results using our method. First, our method can successfully incorporate the gene interactions while other methods ignore this information. Second, our method is capable of dealing with a broad range of primarily affected genes while other methods' performance deteriorates as this number grows. In real perturbation experiments, often multiple genes are primarily affected. Thus, we conclude that our method is more suitable for real perturbation experiments.

3.4. Sensitivity to the gap between primary and secondary effects

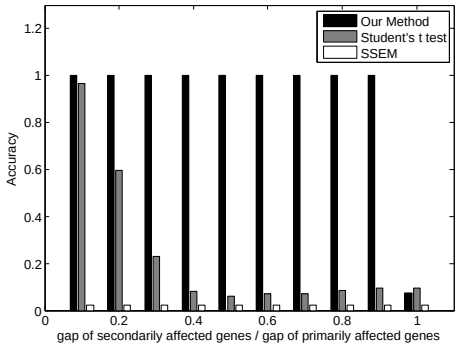


Fig. 5. Comparison of accuracies with SSEM and Student's t test while varying the ratio of gaps of primarily and secondarily affected genes. For a category of gene, the gap denotes the absolute difference of average expressions in control and non-control groups. The x-axis represents the ratio of gaps of primarily and secondarily affected genes. The y axis denotes the accuracy of our method as described inn Section 3.4 The figure demonstrates that our method obtains very high accuracy except when the ratio equals to zero, i.e the gap is equal for both the primarily and secondarily affected genes.

The experiments over the real dataset suggest the validity of our model. One question however follows from these experiments. How does our method compare when we vary the distinction between primarily and secondarily affected genes in terms of their gap between control and non-control data for both those categories of genes. To answer this question we conducted experiments on synthetic datasets, where we change the differences be-

tween primarily and secondarily affected genes and compare our the accuracy of our method with that of SSEM and student's t test.

Synthetic data generation: We generate the data in the presence of a hypothetical perturbation to simulate the real dataset. The primarily and secondarily affected genes are ascertained in the way described in Section 3.3.

To utilize the real dataset to maximum possible extent, we employ an innovative approach. Let us denote the mean of gene g_i in the control and non-control by μ_i and $\mu_i^{'}$, respectively. We subtract the difference $(\mu_i^{'} - \mu_i)$ from all the expressions in the non-control group of g_i. We repeat this subtraction for all the genes. Once the non-control group is *leveled* to control group, we re-modify the non-control expressions of DE genes. If a gene is primarily DE according to the decided set of genes, we increase or decrease its expression over the data points in non-control group by d_p. Similarly, we modify the expression value by d_s, if the gene is secondarily affected. Here, d_p is greater than d_s.

Results: We created three different sets of data by varying d_p and d_s. For all the datasets the number of primarily affected genes was 40. For every dataset, we used different values of d_p given by $\{0.8, 1.2, 1.6\} \times \sigma$, respectively. However, within a dataset d_p was fixed and d_s/d_p ratio was varied as $\{0.1, 0.2, \cdots 1.0\}$. We discuss only the result for the dataset $d_p = 0.8 \times \sigma$ as the results for the other are similar. The accuracy of the methods can fluctuate for different set of affected genes. Hence, for a particular value of d_s and d_p we repeated the experiment five times with different sets of affected genes and averaged the result.

We run the three methods on all the datasets and extract ranks of genes as described in Section 3.3. A higher position in the rank indicates a higher chance of being primarily differentially expressed. Let the set of true primarily affected genes be PA. Let RG be the set of first $|PA|$ genes from the rank produced by a method, where $|PA|$ is the cardinality of PA. We define accuracy of that method as $\frac{|PA \cap RG|}{|RG|}$.

Figure 5 depicts the result from this experiment. It is clear that our method outperforms SSEM all the time. The accuracy of our method is substantially better than Student's t test for all the cases except when the ratio d_s/d_p equals to one. From this experiment, we can conclude that our method performs very well over a wide range of difference between the non-control groups for primarily and secondarily affected genes. Specifically, for the case where these groups have the same mean, our method perform almost as well as the other methods.

4. Conclusion

In this paper, we considered the problem of identifying primarily affected genes in the presence of an external effect that can perturb the expressions of genes. We assumed that we were given the expression measurements of a set of genes before and after the application of an external perturbation. We developed a new probabilistic method to quantify the cause of differential expression of each gene. Our method considers the possible gene interactions in regulatory and signaling networks, for a large number of perturbations. It uses a Bayesian model with the help of Markov Random Fields to capture the dependency between the genes. It also provides the underlying distribution of the impact with confidence interval.

Our experiments on both real and synthetic datasets demonstrated that our method could find primarily affected genes with high accuracy. It achieved significantly better accuracy than two competing methods, namely SSEM and the student's t test method.

Our method produces a probability distribution rather than a fixed binary decision. The major advantage of this approach is that it augment every decision with a range, and hence endows it with a confidence. A distribution is most of the time more useful, as is it models the very stochastic nature of gene interactions.

Acknowledgments

This work was supported partially by NSF under grants CCF-0829867 and IIS-0845439.

References

1. SA. Amundson, M. Bittner, and Y. Chen et al. Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. *Oncogene*, 18(24):3666–72, 1999.
2. Ferhat Ay, Fei Xu, and Tamer Kahveci. Scalable steady state analysis of boolean biological regulatory networks. *PloS one*, 4(12), 2009.
3. KA. Baggerly, KR. Coombes, and KR. Hess et al. Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol*, 8(6):639–59, 2001.
4. Julian Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3):616–618, 1977.
5. Julian Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

6. Christopher M. Bishop. *Pattern Recognition and Machine Learning* . Springer, 2007.
7. RY. Cheng, A. Zhao, and WG. Alvord et al. Gene expression dose-response changes in microarrays after exposure of human peripheral lung epithelial cells to nickel(II). *Toxicol Appl Pharmacol*, 191(1):22–39, 2003.
8. EJ. Cosgrove, Y. Zhou, TS. Gardner, and ED. Kolaczyk. Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics*, 24(21):2482–90, 2008.
9. J. Dausset and Others. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, 6:575–577, 1990.
10. F. Demichelis, P. Magni, and P. Piergiorgi et al. A hierarchical Nave Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC Bioinformatics*, 7:514, 2006.
11. D. di Bernardo, MJ. Thompson, and TS. Gardner et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol*, 23(3):377–83, 2005.
12. S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematics: Berkley*, pages 1496–1517, 1987.
13. G. Giaever, P. Flaherty, and J. Kumm et al. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci U S A*, 101(3):793–8, 2004.
14. G. Giaever, DD. Shoemaker, and TW. Jones et al. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat Genet*, 21(3):278–83, 1999.
15. D. Hamelinck, H. Zhou, and L. Li et al. Optimized normalization for antibody microarrays and application to serum-protein profiling. *Mol Cell Proteomics*, 4(6):773–84, 2005.
16. JM. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.
17. TR . Hughes, MJ. Marton, and AR. Jones et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.
18. T. Ideker, V. Thorsson, and JA. Ranish et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–34, 2001.
19. T. Imaoka, S. Yamashita, and M. Nishimura et al. Gene expression profiling distinguishes between spontaneous and radiation-induced rat mammary carcinomas. *J Radiat Res (Tokyo)*, 49(4):349–60, 2008.
20. B. Jakob, M. Scholz, and G. Taucher-Scholz. Immediate localized CDKN1A (p21) radiation response after damage produced by heavy-ion tracks. *Radiat Res*, 154(4):398–405, 2000.
21. KY. Jen and VG. Cheung. Transcriptional response of lymphoblastoid cells to ionizing radiation. *Genome Res*, 13(9):2092–100, 2003.
22. M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.
23. CM. Kendziorski, MA. Newton, and H. Lan et al. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med*, 22(24):3899–914, 2003.
24. Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer, 2009.
25. R. Lin, Y. Sun, and C. Li et al. Identification of differentially expressed genes in human lymphoblastoid cells exposed to irradiation and suppression of radiation-induced apoptosis with antisense oligonucleotides against caspase-4. *Oligonucleotides*, 17(3):314–26, 2007.
26. PY. Lum, CD. Armour, and SB. Stepaniants et al. Discovering Modes of Action for Therapeutic Compounds Using a Genome-Wide Screen of Yeast Heterozygotes. *Cell*, 116(1):5–7, 2004.
27. MJ. Marton, JL. DeRisi, and HA. Bennett et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med*, 4(11):1293–301, 1998.
28. GL. Miklos and R. Maleszka. Microarray reality checks in the context of a complex disease. *Nat Biotechnol*, 22(5):615–21, 2004.
29. ID. Nagtegaal, CG. Gaspar, and LT. Peltenburg et al. Radiation induces different changes in expression profiles of normal rectal tissue compared with rectal carcinoma. *Virchows Arch*, 446(2):127–35, 2005.
30. AB. Parsons, RL. Brost, and H. Ding et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol*, 22(1):62–9, 2004.
31. KE. Rieger and G. Chu. Portrait of transcriptional responses to ultraviolet and ionizing radiation in human cells. *Nucleic Acids Res*, 32(16):4786–803, 2004.
32. Yishai Shimoni, Gilgi Friedlander, and Guy Hetzroni et al. Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol Syst Biol.*, 3:138, 2007.
33. DA. Smirnov, M. Morley, and E. Shin et al. Genetic analysis of radiation-induced changes in human gene expression. *Nature*, 459(7246):587–91, 2009.
34. Bin Song, I. Esra Buyuktahtakin, Sanjay Ranka, and Tamer Kahveci. Manipulating the steady state of metabolic pathways . *IEEE TCBB*, to appear.
35. R. Storn and K. Price. Differential Evolution A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.
36. KK. Tsai, EY. Chuang, JB. Little, and ZM. Yuan. Cellular mechanisms for low-dose ionizing radiation-induced perturbation of the breast tissue microenvironment. *Cancer Res*, 65(15):6734–44, 2005.
37. CJ. Vaske, C. House, and T. Luu et al. A factor graph nested effects model to identify networks from genetic perturbations. *PLoS Comput Biol*, 5(1):e1000274, 2009.
38. L. Wu and AJ. Levine. Differential regulation of the p21/WAF-1 and mdm2 genes after high-dose UV irradiation: p53-dependent and p53-independent regulation of the mdm2 gene. *Mol Med*, 3(7):441–51, 1997.
39. G. Zhang, CN. Njauw, JM. Park, C. Naruse, M. Asano, and H. Tsao. EphA2 is an essential mediator of UV radiation-induced apoptosis. *Cancer Res*, 68(6):1691–6, 2008.