Identifying Differentially Regulated Genes

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

***Abstract*—Microarray experiments often measure expressions of genes taken from sample tissues in the presence of external perturbations such as medication, radiation, or disease. Typically in such experiments, gene expressions are measured before and after the application of external perturbation. In this paper, we focus on an important class of such microarray experiments that inherently have two groups of tissue samples. The external perturbation can change the expressions of some genes directly or indirectly through gene interaction network. When such different groups exist, the expressions of genes after the perturbation can be different between the two groups. It is not only important to identify the genes that respond differently across the two groups, but also to mine the reason behind this differential response. In this paper, we aim to identify the cause of this differential behavior of genes, whether because of the perturbation or due to interactions with other genes in two group perturbation experiments.**

We propose a new probabilistic Bayesian method with Markov Random Field to find such genes. Our method incorporates information about relationship from gene networks as prior information. Experimental results on synthetic and real datasets demonstrate the superiority of our method compared to existing techniques.

***Keywords*-Gene networks; differentially expressed genes; Markov Random Field;**

I. Introduction

Microarray experiments often measure expressions of genes taken from sample tissues in the presence of external perturbations such as medication, radiation, or disease [1]. Typically in such experiments, gene expressions are measured before and after the application of external perturbation, and are called *control data* and *non-control data*, respectively. In this paper, we focus on an important class of such microarray experiments that inherently have two groups of tissue samples. Different groups in a microarray measurement can exist in many different ways. For instance, samples can be taken from members of multiple closely related species (e.g. rat versus mouse). Within the same species there can be subgroups with different phenotypes (e.g. African American versus Caucasian American). Another example is when the samples have already been through several alternative external perturbations (e.g. fasting or not fasting). When such different groups exist, it is not only important to observe overall changes in gene expression, but also to observe how different groups respond to the external perturbation. For example, Taylor et al. applied medications on 36 Caucasian American (CA) and 33 African American (AA) patients infected with Hepatitis C [2]. Gene expressions were collected before and after the medication.

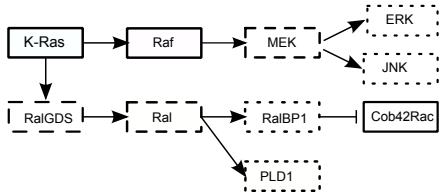


Figure 1. Illustration of the impact of a hypothetical external perturbation on a small portion of the Pancreatic Cancer pathway. The pathway is taken from the KEGG database. The solid rectangles denote the genes affected directly by perturbation, the dashed rectangles indicate genes secondarily affected through the networks. The dotted rectangles denote the genes without any change in expression. $\rightarrow$ 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.

In a perturbation experiment, some of the genes respond by noticeably changing their expression values between the control and non-control data. Genes that change their expression in a statistically significant way are referred to as *differentially expressed (DE)*, while those that do not are referred to as *equally expressed (EE)* genes. In the context of two groups, we refer to a gene that has the same state in both groups, i.e. the gene is DE for both groups or EE for both groups, as *equally regulated (ER)* gene. On the contrary, if a gene is DE in one group and EE in the other, we denote it as *differentially regulated (DR)*.

Genes for any organism typically interact with each other via regulatory and signaling networks. For simplicity, we will refer to them as *gene networks* for the rest of this paper. A small portion of an example gene network can be seen in Figure 1.

Once an external perturbation is applied, a gene can become DE in one of two ways – as a direct effect of the perturbation or via interaction with other DE genes through gene networks. We denote a gene as *primarily affected* DE, if it was due to the external perturbation. Similarly, a gene is *secondarily affected DE*, if it is DE due to another gene in the gene network. Figure 1 shows the state of the genes in the Pancreatic Cancer pathway after a hypothetical external perturbation is applied. In this figure, genes K-Ras, Raf and Cob42Roc are primarily affected and MEK, Ral and

RalGDS are secondarily affected through interactions.

Recall that for a gene to be DR, it has to be EE in one group and DE in another group. For such a gene, if it happens to be DE in one group because of the external perturbation, we call it as *primarily differentially regulated (PDR)* gene. When it is DE in one group because of the interaction with other DE genes in the gene networks, we will refer to it by *secondarily differentially regulated (SDR)* gene. *In this paper, we consider the problem of identifying the PDR genes in a given set of control and non-control gene expressions from two groups of samples*

Existing methods estimate differentially expressed genes in multiple groups of tissue samples [3], [4], though they do not specify the cause of differential regulation for those genes such as primary or secondary. There are also some methods to analyze the primary and secondary effects when there is only a single group [5], [6]. A simple solution to find the PDR genes is to apply this method on each group separately and find the difference between the two sets. Our experimental results demonstrate that this approach has low accuracy. We discuss this in more detail in Section IV.

Our approach: In this paper, we propose a new probabilistic Bayesian method to find the PDR genes in microarray dataset as defined above. Our Bayesian model based method incorporates information about relationship from gene networks as a prior belief. We consider the gene network as a directed graph where each node represents a gene, and a directed edge from gene g_i to gene g_j represents a genetic interaction (e.g g_i activates or inhibits g_j). We define two genes as *neighbors* of each other if they share a directed edge. We also classify a neighbor as *incoming* or *outgoing*, if it is at the start or at the end of the directed edge respectively. When the expression level of a gene is altered, it can affect some of its outgoing neighbors. Thus, the gene expression can change due to external perturbation or because of one or more of the affected incoming neighbors.

We represent the external perturbation by a hypothetical gene (i.e. *metagene*) g_0 in the gene network. We add an edge from the metagene to all the other genes because 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 DR due to an alteration in the activity of gene g_i ($\forall$ g_i, $g_j \in \mathcal{G} \cup \{g_0\}$) 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 define feature functions that encapsulate the domain knowledge available from gene networks and gene expression data; and optimize the joint posterior distribution over the random variables in the MRF using Iterated Conditional Mode (ICM) [7]. The optimization provides the state of the genes, the regulation of the genes and the probabilistic estimate of pairwise interactions between the genes including the metagene. Given this, we can rank the genes according to the probability that a gene is DR because of the metagene g_0, and obtain a list of possible PDR genes.

We compare the accuracy of our method with SSEM and Student's t-test [8] on semi-synthetic dataset generated from microarray data in Cosgrove et al [9]. We also compare against a method to identify the primarily affected DE genes in a single group perturbation data developed by Bandyopadhyay et al. [5]. Our method obtains high accuracy and outperforms all the other three methods.

The rest of the paper is organized as follows. Section II provides an overview of the related work. Section III discusses our method. Section IV presents the experimental results. Section V concludes the paper.

II. Related Work

Existing methods to identify the primarily affected DE genes using association analysis techniques [10], haploinsufficiency profiling [11] and chemical-genetic interaction mapping [12] are limited to applications where additional information such as fitness based assays of drug response or a library of genetic mutants is available. Bernardo et al. suggested a regression based approach named MNI that assumes that the internal genetic interactions are offset by the external perturbation [6]. It estimates gene-gene interaction coefficients from the control data and uses them to predict the target genes in the non-control data. Cosgrove et. al. proposed a method named SSEM that is similar to MNI [9]. SSEM models the effect of perturbation by an explicit change of gene expression from that of the unperturbed state.

In our previous work, we have developed a method to detect the primarily and secondarily affected genes in perturbation experiments with a single group [5]. We will call this method SMRF (single MRF) in the rest of this paper for it applies MRF on single group datasets. A Bayesian probabilistic method based on Markov Random Field is developed that leverages the information from gene networks as the prior belief of the model. Though these methods analyze primary and secondary effects of perturbation on gene expressions, they are not directly applicable for multi-group perturbation experiments.

Several recent studies aim to identify DE genes in multiple groups of data points. maSigPro is a two stage regression based method that identifies genes that demonstrate differential gene expression profiles across multiple experimental groups [3]. Hong et al. proposed a functional hierarchical model for detecting temporally differentially expressed genes between two experimental conditions [13]. They model gene expressions by basis function expansion and estimate the parameters using a Monte Carlo EM algorithm. Tai et al. ranked DE genes using data from replicated microarray time course experiments, where there

are multiple biological conditions [4]. They derived a multi-sample multivariate empirical Bayes statistic for ranking genes. Angelini et al. proposed a Bayesian method for detecting temporally DE genes between two experimental conditions [14]. Deun et al. develops a Bayesian method to find the genes that are differentially expressed in a single tissues or condition over multiple tissues or conditions [15]. All these methods identify differentially expressed genes in multiple groups. *However, none of these methods analyzes the primary and secondary effects in a two group perturbation experiment. In this paper, we develop a method to solve this problem.*

III. Methods

In this section we describe different aspects of our method. Section III-A describes the relevant notation and formulates the problem. Section III-B provides a high level overview of the solution. Section III-C describes the calculation of the prior density function of MRF.

A. Notation and problem formulation

In this section, we describe our notation and formally define the problem. We classify the random variables of the model into two different groups, namely *observed variables* and *hidden variables*.

Observed variables: We define two sets of observed variables, one for microarray gene expression data and another for the neighborhood in the extended gene network.

- **Microarray data:** We denote the number of genes by M and the number of data points in the two groups D_A and D_B by N_A and N_B respectively. We represent the set of genes with $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$. We refer to the complete gene expression data using variable $\mathcal{Y}$.
- **Neighborhood variables:** We use the term $\mathcal{W} = \{W_{ij}\}$ to indicate if any two genes g_i and g_j are neighbors in the extended gene network. If g_i is an incoming neighbor of g_j (i.e. g_j has an incoming edge from g_i), then we set the value of W_{ij} $(1 \leq i, j \leq M)$ to 1. It is 0 otherwise.

Hidden variables: We define three sets of hidden variables. These variables govern the state of genes, regulations of genes and interactions among genes respectively.

- **State variables:** We use $\mathcal{S}_A = \{S_{Ai}\}$ and $\mathcal{S}_B = \{S_{Bi}\}$, $(1 \leq i \leq M)$ to denote the states of the genes in group D_A and D_B. $S_{Ai} = 1$ if g_i is DE in D_A and 0 if it is EE in D_A. We define S_{Bi} similarly. We assume that the metagene g_0 is DE for both D_A and D_B. Thus, $S_{A0} = S_{B0} = 1$.
- **Regulation variables:** We denote the regulation condition of gene g_i with Z_i. Table I enumerates different values of Z_i for the values of S_{Ai} and S_{Bi}. In this formulation, the cases $Z_i = \{2, 3\}$ indicate that g_i is DR, whereas $Z_i = \{1, 4\}$ indicate that g_i is ER. The metagene is guaranteed to be ER, since $S_{A0} = S_{B0} = 1$.
- **Interaction variables:** In order to govern the joint regulation states of genes g_i and g_j we define interaction variables $\mathcal{X} = \{X_{ij}\}, (1 \leq i, j \leq M)$. Mathematically, $X_{ij} = 4 \times (Z_i - 1) + Z_j$. Table I enumerates different values of X_{ij} for values of Z_i and Z_j. Specifically, $X_{0j} \in \{2, 3\}$ and $X_{0j} \in \{1, 4\}$ correspond to the cases where g_j is DR and ER respectively because of interaction with the metagene g_0.

Table I
Enumeration of the values of Z_i, Z_j and X_{ij} for different values of S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj}. The hidden variables are oriented in a hierarchical structure. For instance, the value of Z_i depends on the values of S_{Ai} and S_{Bi}. Similarly, the value of X_{ij} depends on the values of Z_i and Z_j. Thus, the value of the dependent variable X_{ij} in tern depends on the values of four independent variables S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj}.

S_{Ai}	S_{Bi}	S_{Aj}	S_{Bj}	Z_i	Z_j	X_{ij}
DE	DE	DE	DE	1	1	1
DE	DE	DE	EE	1	2	2
DE	DE	EE	DE	1	3	3
DE	DE	EE	EE	1	4	4
DE	EE	DE	DE	2	1	5
DE	EE	DE	EE	2	2	6
DE	EE	EE	DE	2	3	7
DE	EE	EE	EE	2	4	8
EE	DE	DE	DE	3	1	9
EE	DE	DE	EE	3	2	10
EE	DE	EE	DE	3	3	11
EE	DE	EE	EE	3	4	12
EE	EE	DE	DE	4	1	13
EE	EE	DE	EE	4	2	14
EE	EE	EE	DE	4	3	15
EE	EE	EE	EE	4	4	16

It is easy to see that the hidden variables follow a hierarchical structure. For instance, the value of Z_i depends on the values of S_{Ai} and S_{Bi}. Similarly, the value of X_{ij} depends on the values of Z_i and Z_j. Thus, the value of the dependent variable X_{ij} depends on the values of four independent variables S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj}. Table I enumerates the values of Z_i, Z_j and X_{ij} for different values of S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj}.

Problem formulation: Let $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$ denote the set of all genes. Using the definitions of the neighborhood variables $\mathcal{W}$, we denote the collection ($\mathcal{G}$, $\mathcal{W}$) by $\mathcal{V}$ which essentially represents the gene networks. We denote the metagene by g_0. Given an observed data $\{\mathcal{V}, \mathcal{Y}\}$ we want to estimate the probabilities $p(X_{ij} = x | \mathcal{X} - X_{ij}, \mathcal{Y}, \mathcal{V})$, $x \in \{1, 2, \cdots 16\}$.

A higher value of $p(X_{0j} = \{2, 3\}|\cdot)$ indicates a higher probability of a gene g_j being PDR. Using the estimated values of $p(X_{0j}|\cdot), \forall j \in \{1, 2, \cdots M\}$, we can create an ordered list of candidate PDR genes.

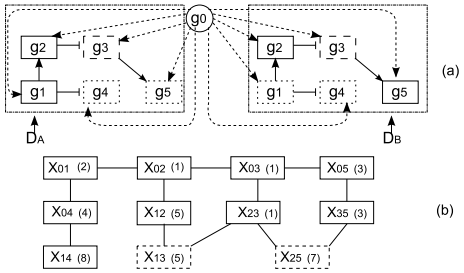


Figure 2. **(a)**. A small hypothetical gene network with perturbation in two datasets D_A and D_B. The genes in the two datasets interact through identical network, although they assume different states. The circle g_0 represents the abstraction of the external perturbation. Rectangles denote genes. $\rightarrow$ implies activation and $\dashv$ implies inhibition. The potential effect of metagene to all other genes is indicated by dotted arrows from the metagene to all the other genes. For example, g_1 is primarily affected in D_A, but not affected in D_B. g_2 is primarily affected in both the datasets. g_3 is secondarily affected in both D_A and D_B. (b) The Markov Random Filed graph constructed based on the small hypothetical gene network in (a). The numbers in the parenthesis are the expected assignments to the variables based on the states of the genes in (a). Nodes with dotted boundaries indicate that those nodes are required for completeness of the model, however the corresponding interactions do not exist.

B. Overview of the solution

We build a Bayesian probabilistic method based on Markov Random Field where we leverage the information from gene networks as the prior belief of the model. Using Bayes theorem [16] we can write the joint probability density of interaction variables $\mathcal{X}$ as

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

The first term in the numerator, $P(\mathcal{Y}|\mathcal{X},\mathcal{V},\theta_Y)$, is the likelihood of the observed expression data $\mathcal{Y}$ given the interaction variables and gene network. θ_Y represents the parameters for the likelihood function.

The second term in the numerator $P(\mathcal{X}|\mathcal{V},\theta_X)$, represents this prior belief. θ_X represents the parameters for the prior density function. We define a Markov Random Field (MRF) over the interaction variables $\mathcal{X}$ and the priors are encoded via feature functions in the MRF. Details of the priors and the associated feature functions are outlined in Section III-C. The denominator of Equation 1 is the normalization constant that represents the sum of the product of the likelihood and the prior over all possible assignments of interaction variables $\mathcal{X}$.

Given the joint probability density function outlined in Equation 1, our original problem reduces to obtaining assignments for the interaction variables $\mathcal{X}$ and the parameters θ_X and θ_Y that maximize it.

We use a pseudo-likelihood version of the objective function to optimize the objective. We use Iterative Conditional Mode (ICM) [7] and Differential Evolution [17] in an alternating optimization technique to maximize the pseudo-likelihood with respect to $\mathcal{X}$, θ_X and θ_Y.

After the optimization, we obtain an assignment for $\mathcal{X}$, θ_X and θ_Y. Using these assignments and the observed data, we estimate the posterior probability of all X_{ij} variables. Using the estimated values of $p(X_{0j}|\cdot), \forall j \in \{1, 2, \cdots M\}$, we create an ordered list of candidate PDR genes. We elaborate on the derivation of prior density function in the next section.

C. Computation of the prior density function

In this section, we describe the way we incorporate gene network, as the the prior belief into our Bayesian model. From the structure and properties of gene network, we build three hypotheses and embed them into our model.

- In each group D_T ($T \in \{A, B\}$), the metagene g_0 can change the state of all the other genes.
- In each group D_T ($T \in \{A, B\}$), a gene g_i can change the states of its outgoing neighbors g_j in the same data group.
- Each gene has a high probability of being equally regulated. This follows from the observation that, often the difference between the expressions of most of the genes in two groups is small.

In order to compute the prior density function, we define a Markov Random Field (MRF) over the $\mathcal{X}$ variables. 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 and the two genes corresponding to the other subscript interact in the 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). We formulate these dependency constraints of the nodes in MRF using a set of unary and binary functions called *feature functions*. We discuss these feature functions next.

We denote the neighbors of X_{ij} in the MRF graph as X^*_{ij} = $\{X_{kj}|W_{ki} = 1\} \cup \{X_{ip}|W_{jp} = 1\}$. 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. 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, $p(\mathcal{X}|\theta_X) = \frac{1}{\Delta}\prod_{C_{ij}, W_{ij}=1} \psi(C_{ij})$ [18]. In this formulation, Δ is the normalization function $\Delta = \sum_{\mathcal{X}} \prod_{C_{ij}} \psi(X_{ij})$. To limit the complexity of our model, we consider only cliques of size one and two.

We define seven feature functions to capture the dependencies among the variables in $\mathcal{X}$ according to the three

hypothesis.

Unary feature functions F_1, F_2, F_3: A primary component of the prior density function is modeling the frequency of X_{ij} itself. Here, we focus to on two values of X_{ij} namely X_{ij} = {2, 3}, since they correspond to the events that a gene g_j is DR due to the metagene g_0. When X_{ij} = 2, g_j is DE in D_A and EE in D_B. To capture this, we define a feature function $F_1(X_{ij})$ which returns one when X_{ij} = 2 and zero otherwise. Similarly, X_{ij} = 3 when g_j is EE in D_A and DE in D_B. We define another feature function $F_2(X_{ij})$, which returns one when X_{ij} = 3. We capture all the other values of X_{ij} by a feature function called $F_3(X_{ij})$.

Binary feature functions F_4, F_5: Let Υ represent the hypothesis that in a group $D_T, T \in \{A, B\}$ a gene g_j including the metagene can change the state of one of its outgoing neighbors g_k. We make a stronger hypothesis Υ^o that, Υ holds simultaneously in D_A and D_B with high probability. Note that, this stronger hypothesis is based on the assumption that the genes in both D_A and D_B express in a similar fashion. This assumption is meaningful as in these two-group perturbation experiments the different groups belong to similar biological conditions [2].

Υ^o is encoded in $\mathcal{X}$ domain as follows. Consider four genes g_p, g_i, g_j and g_k, such that $g_p \rightarrow g_i$, $g_i \rightarrow g_j$ and $g_j \rightarrow g_k$. Here $\rightarrow$ indicates that the gene on the left activates or inhibits the gene on the right. By definition, (X_{pj}, X_{ij}) and (X_{ij}, X_{ik}) are edges in the MRF. Note that the first edge corresponds to an incoming neighbor g_p of g_i, while the second edge corresponds to an outgoing neighbor g_k of g_j. We discriminate between these two sets of neighbors of X_{ij}, as they are related to the incoming neighbors of g_i and outgoing neighbors of g_j respectively. It can be shown that, for the first set of edges, X_{pj} equals to X_{ij} if and only if (*iff*) $Z_p = Z_i$, i.e. Υ^o holds true. Similarly, for the second set of edges X_{ij} equals to X_{ik} *iff* $Z_j = Z_k$, which in turn implies that Υ^o is satisfied.

We define two sets of feature functions to formalize these equalities based on the incoming neighbors of g_i and the outgoing neighbors of g_j.

- **Left external equality:** We denote the incoming neighbors of g_i with $In(g_i)$. We write a feature function $f_4(X_{pj}, X_{ij})$, $\forall p$, $g_p \in In(g_i)$. $f_4(X_{pj}, X_{ij})$ = 1 if $Z_i = Z_p$ and $W_{pi} = W_{ij} = 1$. Otherwise, $f_4(X_{pj}, X_{ij})$ = 0. We denote the summation of this function over all the incoming neighbors of g_i as $F_4(X_{ij}) = \sum_{p,W_{ij}=1,W_{pi}=1} f_4(X_{ij}, X_{pj})$.
- **Right external equality:** We denote the outgoing neighbors of g_j as $Out(g_j)$. We define a feature function $f_5(X_{ik}, X_{ij})$, $\forall k$, $g_k \in Out(g_j)$. $f_5(X_{ik}, X_{ij})$ = 1 if $S_k = S_j$ and $W_{jk} = W_{ij} = 1$. Otherwise, $f_5(X_{ik}, X_{ij})$ = 0. We denote the summation of this function over all the outgoing neighbors of g_j as $F_5(X_{ij}) = \sum_{k,W_{ij}=1,W_{jk}=1} f_5(X_{ij}, X_{ik})$.

Unary feature functions F_6, F_7: We introduce two unary feature functions to incorporate our last hypothesis, that all genes are ER with a high probability. We consider two genes g_i and g_j such that $g_i \rightarrow g_j$. This hypothesis holds true, if g_i is equally regulated or g_j is equally regulated.

- **Left internal equality :** We define this feature function to capture the events when g_i is equally regulated. As, g_j can assume any state, this feature function holds true for eight different values of X_{ij}. We denote the feature function by $f_6(X_{ij}, t)$ that returns one if its two arguments are equal and zero otherwise. We denote the summation of this functions over all these eight values of X_{ij} as $F_6(X_{ij}) = \sum_{i,j,W_{ij}=1,t\in\{1,\cdots,4,13,\cdots,16\}} f_6(X_{ij}, t)$.
- **Right internal equality:** We define this feature function to capture the events when g_j is equally regulated. As, g_i can assume any state, this feature function holds true for eight different values of X_{ij}. We denote the feature function by $f_7(X_{ij}, t)$ that returns one if its two arguments are equal and zero otherwise. We denote the summation of this functions over all these eight values of X_{ij} as $F_7(X_{ij}) = \sum_{i,j,W_{ij}=1,t\in\{5,\cdots,12\}} f_7(X_{ij}, t)$.

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

$$p(\mathcal{X}|\theta_X) = \frac{1}{\Delta} exp\Big(\sum_{i,j,W_{ij}=1,k\in\{1,2,\cdots,7\}} \gamma_k F_k(X_{ij}) \Big) \quad (2)$$

In the above equation γ_k, $k \in \{1, 2, \cdots 7\}$ are the coefficients of the seven feature functions in MRF.

IV. 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 the code for Differential Evolution from http://www.icsi.berkeley.edu/~storn/code.html. We compared our method with SSEM as SSEM is one of the most recent methods that considers identifying primarily affected genes in a perturbation experiments [9].

We obtained SSEM from http://gardnerlab.bu.edu/SSEMLasso. We executed our program on a Quad-Core AMD Opteron 2 Ghz workstation with 32 GB of memory. We compared the accuracy of our method to three such methods namely, SMRF [5], Student's t-test and SSEM.

Dataset: We used two different sets of data to conduct the experiments in this paper. We describe two of them here. The first dataset (dataset 1) was collected by Smirnov et al. [19]. 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 [20]. Microarray snapshots were obtained at 0th hour (i.e., before the radiation) and 2 and 6 hours after the radiation.

For this dataset, we adapted the time series data to create the control and non-control data for our experiments. We used the data before perturbation as control data. For the non-control data we calculated the expected expressions of a gene at each points after the perturbation. We selected the one with highest absolute difference from the expected expression of control data for that gene.

We created the second dataset from dataset 1 using the sigmoid method described in Garg et al. [21]. The sigmoid method computes the gene expressions change over time when the current expressions are provided. We used the control group of dataset 1 as the control group of dataset 2. Then, we changed the expression values of some of the randomly selected genes to model the primary effect of external perturbation. We applied the sigmoid method on the resulting dataset to create dataset 2.

We also collected 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 considered only the genes that take part in the gene networks in our model.

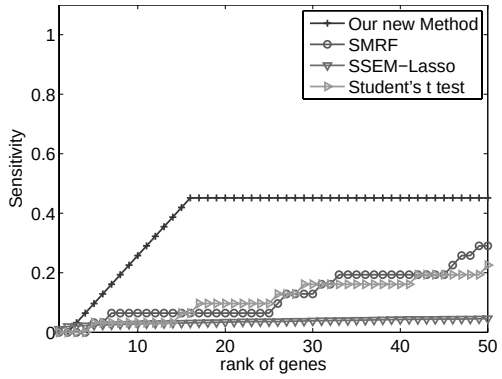
A. Comparison to other methods

Our method provides us a list of differentially regulated genes. We sort the list of those genes as follows. Consider a DR gene g_i, which is DE in D_A and EE in D_B. We calculate the likelihood of being EE in D_A and DE in D_B for that gene. We can interpret it as the probability of being DR but in a reverse way. We could instead use the probability that the gene is DE in D_A and EE in D_B. However, according to our observation, the earlier metric provides a much better accuracy. We sort all the DR genes with increasing order of that likelihood.

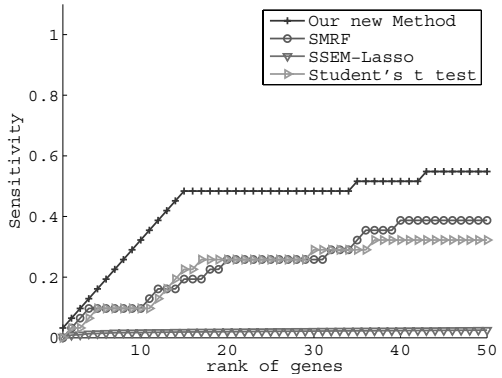
As per our knowledge, no other method exists that differentiates between the primary and secondary effects in a two-group perturbation experiment. There exist some studies in identifying primarily affected genes in single group datasets.

Experimental setup: Given an input dataset, using each of the four methods, we ranked all the genes. Highly ranked genes have higher chance of being a primarily differentially regulated according to each method. However, as other three methods are not tailored to solve this problem, we created separate ranking on D_A and D_B. Then, out of those two lists, we created a unified rank of differentially regulated genes. We shall elaborate on this unified rank creation later. In the complete version, we explain how we create ranks on individual groups D_A and D_B for other three methods.

Now we describe how we create an unified ranking system of differentially regulated genes for these three methods. We denote the ranks from data group D_A and D_B by R_A and R_B respectively. The unified rank is defined by R_U. We denote the number of genes in each rank to be ω_A and ω_B respectively. We scan both the ranks simultaneously from first position to ω = min(ω_A, ω_B). While scanning at the



(a) deviation = 0.6



(b) deviation = 0.8

Figure 3. Comparison of our method to SMRF, SSEM and t-test. The number of primarily differentially affected genes is 40. The gaps between the mean of control and non-control groups in primarily affected genes are 0.02 and 0.1 $\times\sigma$, where sigma is estimated from the real dataset. The figures indicate that our method outperforms SMRF, SSEM and t-test.

kth position, we denote the equally regulated set obtained till that position by $\Lambda_k = R_A(1:k) \cap R_B(1:k)$. We include $R_T(k)$ to the unified rank R_U if $R_T(k) \notin \Lambda_k, , T \in \{A, B\}$. For SSEM we obtain a separate R_U for each data point. We average the accuracies over all those ranks.

Results: In these experiments we used dataset 2, that we have just described. To observe the accuracy of our method at varying degrees of difficulties, we conducted experiments with four different values of deviation, namely, $\{0.5, 0.6, 0.7, 0.8\}$. However, due to space limitation we discuss only two of them in this paper (see Figure 3). The results we discuss correspond to the cases when deviation = $\{0.6, 0.8\}$.

Figures 3(a)and 3(b) show the sensitivity of the four methods when with the two deviation settings. The former one corresponds to the computationally harder case as the difference between the non-control groups of primarily and secondarily affected genes is small. As the deviation increases identifying primarily affected genes becomes easier.

Form the figure, we observe that our method is significantly more accurate than the other three methods for all datasets consistently. It reaches almost 50% sensitivity (i.e.,

it can find around 15-18 primarily affected genes out of 30) in the top 50 ranked genes, when the deviation is 0.6. On the other hand, it achieves a sensitivity of 0.6 when the deviation is 0.8. The results are similar for other deviations (results not shown). The method in SMRF reaches to 30% and 40% accuracy, however in a sower pace. The t-test reaches around 25% and 30% sensitivity at 50 ranking position respectively. SSEM's sensitivity is below 0.1 for all experiments even within the top 50 positions.

V. Conclusion

We solved the problem of finding primarily differentially regulated genes in the presence of external perturbations and two groups. The probabilistic Bayesian method with Markov Random Field incorporates information about relationship from gene networks as prior information. Experimental results on synthetic and real datasets demonstrated the superiority of our method compared to simple techniques that find primarily differentially expressed genes in two classes and find the difference between them.

Acknowledgment

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

References

[1] R. Cheng, A. Zhao, and W. A. et al., "Gene expression dose-response changes in microarrays after exposure of human peripheral lung epithelial cells to nickel(II)." *Toxicol Appl Pharmacol*, vol. 191, no. 1, pp. 22–39, 2003.

[2] M. Taylor, T. Tsukahara, and L. Brodsky, "Changes in gene expression during pegylated interferon and ribavirin therapy of chronic hepatitis c virus distinguish responders from non-responders to antiviral therapy." *J Virol*, vol. 81, no. 7, pp. 3391–401, 2007.

[3] A. Conesa, M. Nueda, A. Ferrer, and M. Talin, "masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments." *Bioinformatics*, vol. 22, no. 9, pp. 1096–102, 2006.

[4] Y. Tai and T. Speed, "On gene ranking using replicated microarray time course data." *Biometrics*, vol. 65, no. 1, pp. 40–51, 2009.

[5] N. Bandyopadhyay, M. Somaiya, T. Kahveci, and S. Ranka, "Modeling Perturbations using Gene Networks," in *International Conference on Computational Systems Bioinformatics*, 2010.

[6] D. di Bernardo, M. Thompson, and T. G. et al., "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." *Nat Biotechnol*, vol. 23, no. 3, pp. 377–83, 2005.

[7] J. Besag, "On the Statistical Analysis of Dirty Pictures." *Journal of the Royal Statistical Society*, vol. 48, no. 3, pp. 259–302, 1986.

[8] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics: An Introduction (Statistics for Biology and Health)*, 2nd ed. Springer, December 2004.

[9] E. Cosgrove, Y. Zhou, T. Gardner, and E. Kolaczyk, "Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia." *Bioinformatics*, vol. 24, no. 21, pp. 2482–90, 2008.

[10] T. . Hughes, M. Marton, and A. J. et al., "Functional discovery via a compendium of expression profiles." *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[11] G. Giaever, D. Shoemaker, and T. J. et al., "Genomic profiling of drug sensitivities via induced haploinsufficiency." *Nat Genet*, vol. 21, no. 3, pp. 278–83, 1999.

[12] A. Parsons, R. Brost, and H. D. et al., "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." *Nat Biotechnol*, vol. 22, no. 1, pp. 62–9, 2004.

[13] F. Hong and H. Li, "Functional hierarchical models for identifying genes with different time-course expression profiles." *Biometrics*, vol. 62, no. 2, pp. 534–44, 2006.

[14] C. Angelini, D. De Canditiis, and M. Pensky, "Bayesian models for two-sample time-course microarray experiments," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1547–1565, March 2009.

[15] K. V. Deun, H. Hoijtink, and L. T. et al., "Testing the hypothesis of tissue selectivity: the intersection-union test and a bayesian approach." *Bioinformatics*, vol. 25, no. 19, pp. 2588–94, 2009.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning* . Springer, 2007.

[17] R. Storn and K. Price, "Differential Evolution A Simple and Efficient Heuristic for global Optimization over Continuous Spaces." *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[18] J. Hammersley and P. Clifford, "Markov fields on finite graphs and lattices." *Unpublished manuscript*, 1971.

[19] D. Smirnov, M. Morley, and E. S. et al., "Genetic analysis of radiation-induced changes in human gene expression." *Nature*, vol. 459, no. 7246, pp. 587–91, 2009.

[20] J. Dausset and Others, "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome," *Genomics*, vol. 6, pp. 575–577, 1990.

[21] A. Garg, L. Mendoza, I. Xenarios, and G. De Micheli, "Modeling of Multiple Valued Gene Regulatory Networks," in *29th Annual International Conference of the IEEE EMBS*, vol. IEEE CNS, 2007, pp. 1398 – 1404.

[22] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.