BMC Genomics

RESEARCH Open Access

CMRF: analyzing differential gene regulation in two group perturbation experiments

Nirmalya Bandyopadhyay*, Manas Somaiya, Sanjay Ranka, Tamer Kahveci

From First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011)
Orlando, FL, USA. 3-5 February 2011

Abstract

Background: Microarray experiments often measure expressions of genes taken from sample tissues in the presence of external perturbations such as medication, radiation, or disease. The external perturbation can change the expressions of some genes directly or indirectly through gene interaction network. In this paper, we focus on an important class of such microarray experiments that inherently have two groups of tissue samples. When such different groups exist, the changes in expressions for some of the 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.

Results: We propose a new probabilistic Bayesian method *CMRF* based on Markov Random Field to identify such genes. CMRF leverages the information about gene interactions as the prior of the model. We compare the accuracy of CMRF with SSEM and Student's t test and our old method SMRF on semi-synthetic dataset generated from microarray data. CMRF obtains high accuracy and outperforms all the other three methods. We also conduct a statistical significance test using a parametric noise based experiment to evaluate the accuracy of our method. In this experiment, CMRF generates significant regions of confidence for various parameter settings.

Conclusions: In this paper, we solved the problem of finding primarily differentially regulated genes in the presence of external perturbations when the data is sampled from two groups. The probabilistic Bayesian method CMRF based on Markov Random Field incorporates dependency structure of the gene networks as the prior to the model. Experimental results on synthetic and real datasets demonstrated the superiority of CMRF compared to other simple techniques.

Background

Microarray experiments often measure expressions of genes taken from sample tissues in the presence of external perturbations such as medication, radiation, or disease [1,2]. 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 and 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

* Correspondence: nirmalya@cise.ufl.edu
Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32603, USA

and 33 African American patients infected with Hepatitis C [3]. Gene expressions were collected before and after the medication.

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 expressions 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 the groups, i.e. either DE or EE for both the 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 be DE in one of the 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 is DE 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 to identify the primarily affected DE genes using association analysis techniques [4,5], haploinsufficiency profiling [6-8] and chemical-genetic interaction mapping [9] 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 [10]. 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 [11]. SSEM models the effect of perturbation by an explicit change of gene expression from that of the unperturbed state.

We have also developed a method to detect the primarily and secondarily affected genes in perturbation experiments with a single data group [12]. We will call this method SMRF (single MRF) in the rest of this paper for it applies MRF on single group datasets. In that paper we developed a Bayesian probabilistic method based on Markov Random Field that leverages the information from gene networks as the prior belief of the model.

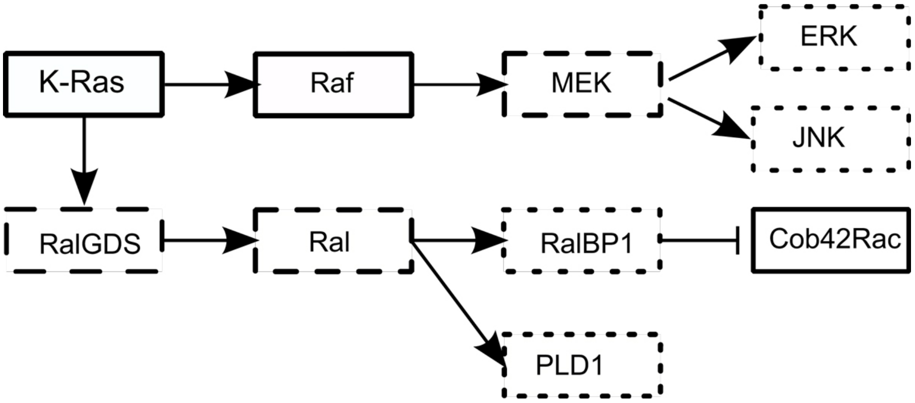


Figure 1 A sample gene regulatory network. Illustration of the impact of a hypothetical external perturbation on a small portion of the Pancreatic Cancer pathway. The pathway is taken from the KEGG database. The solid rectangles denote the genes affected directly by perturbation, the dashed rectangles indicate genes secondarily affected through the networks. The dotted rectangles denote the genes without any change in expression. → implies activation and ⊣ implies inhibition. In this figure, genes K-Ras, Raf and Cob42Roc are primarily affected and MEK, Ral and RalGDS are secondarily affected through interactions.

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 [13]. Hong et al. proposed a functional hierarchical model for detecting temporally differentially expressed genes between two experimental conditions [14]. They modeled 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 [15]. They derived a multisample multivariate empirical Bayes statistic for ranking genes. Angelini et al. proposed a Bayesian method for detecting temporally DE genes between two experimental conditions [16]. Deun et al. developed a Bayesian method to find the genes that are differentially expressed in a single tissue or condition over multiple tissues or conditions [17]. All these methods identify differentially expressed genes in multiple groups. *However, none of these methods analyzed the primary and secondary effects in a two group perturbation experiment. In this paper, we develop a method to solve this problem.*

Our approach

In this paper, we propose a new probabilistic Bayesian method CMRF to find the PDR genes in two group perturbation experiment dataset as defined above. We call this method CMRF (Comparative MRF) for it applies MRF on two groups of data for comparison purpose. Our Bayesian method incorporates information about relationship from gene networks as prior beliefs. 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. For example, in Figure 1, genes K-Ras and Raf are neighbors as K-Raf activates Ras. We also classify a neighbor 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. 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*. CMRF estimates the probability that a gene g_j is DR due to an alteration in the activity of gene $g_i(\forall g_i \in \mathcal{G} \cup \{g_0\},\ g_j \in \mathcal{G})$ 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) [18] 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. CMRF optimizes the joint posterior distribution over the random variables in the MRF using Iterated Conditional Modes (ICM) [19]. 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 data likelihood that a gene is DR because of the metagene g_0, and obtain a list of possible PDR genes.

Figure 2 illustrates different components of CMRF and the connectivity between them. Note that, (C) corresponds to the Bayesian prior based on MRF.

We compare the accuracy of CMRF with that of SSEM and Students t test on semi-synthetic dataset generated from microarray data in Cosgrove et al [11]. We also compare CMRF with our old method SMRF that we developed to identify the primarily affected DE genes in a single group perturbation data [12]. CMRF obtains high accuracy and outperforms all the other three methods. Also, we conduct a statistical significance test using a parametric noise based experiment to evaluate the accuracy of CMRF. In this experiment our model demonstrates reasonable confidence regions for various values of the parameters.

The rest of the paper is organized as follows. Section Results and discussion presents the results of our experiments. Section Methods describes our methods in detail. Section Conclusions concludes our discussion.

Results and discussion

In this section we discuss the experiments we conducted to evaluate the quality of CMRF. We implemented CMRF in MATLAB and Java. We obtained the code for Differential Evolution from http://www.icsi.berkeley.edu/~storn/code.html. We compared CMRF with SSEM as SSEM is one of the most recent methods that considers identifying primarily affected genes in a perturbation experiments [11].

We obtained SSEM from http://gardnerlab.bu.edu/SSEMLasso. We executed our code on a Quad-Core AMD Opteron 2 Ghz workstation with 32 GB of memory.

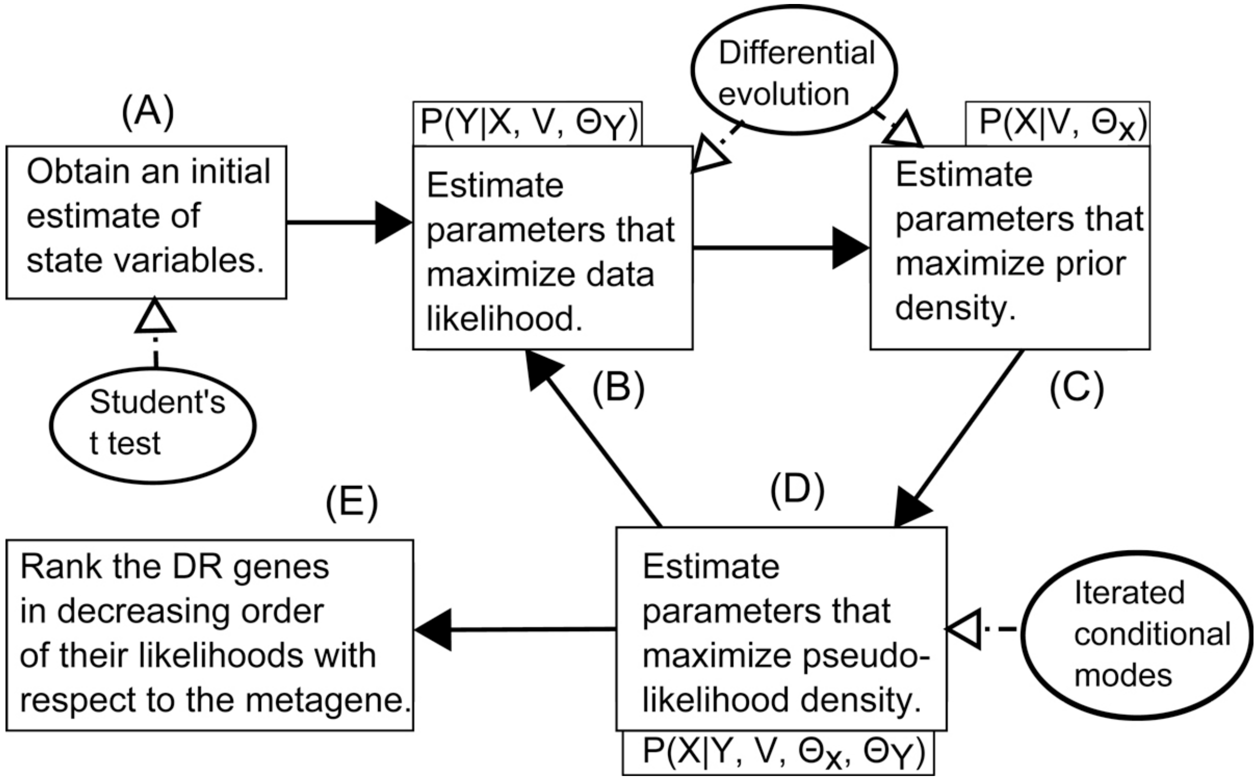

(A)
Obtain an initial estimate of state variables.
Student's t test
P(Y|X, V, ΘY)
Estimate parameters that maximize data likelihood.
Differential evolution
P(X|V, ΘX)
Estimate parameters that maximize prior density.
(B)
(C)
(D)
(E)
Rank the DR genes in decreasing order of their likelihoods with respect to the metagene.
Estimate parameters that maximize pseudo-likelihood density.
Iterated conditional modes
P(X|Y, V, ΘX, ΘY)

Figure 2 Illustration of different components of CMRF and connectivity between them. (A) obtains an initial estimates of state variables using Student's t test. (B) estimates parameters in a way that maximizes data likelihood. (C) estimates parameters in order to maximize prior density. Both (B) and (C) use a global optimization technique called *Differential Evolution*. (D) employs *Iterated Conditional Modes* to maximize the pseudo-likelihood. (B), (C) and (D) consist of an alternating optimization technique. These three steps (B), (C) and (D) are repeated till the algorithm meets a criteria for completion. Finally, once the optimization is complete, the DR genes are sorted in decreasing order of their likelihood with respect to the metagene g_0. The genes at the top of the list are declared PDR.

Dataset

We used four different sets of data to conduct the experiments in this paper.

- Dataset 1. The first dataset was collected by Smirnov et al. [20]. This 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 [21]. Microarray snapshots were obtained before (at zero*th* hour) and after (at second and sixth hours) the application of radiation.
- Dataset 2. The second dataset corresponds to a drug response experiment conducted by Taylor et al [3]. Medications were applied on 36 Caucasian American and 33 African American patients infected with Hepatitis C. Gene expressions were collected before the medication was started and at 1, 2, 7, 14, 28 days after the medication was administered.

Both dataset 1 and 2 are microarray time series data with more than two time points. We adapted these two time series data two create control and non-control data suitable for our experiments. We used the data before perturbation as control data. For the non-control data we calculated the expected expression 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.

- Dataset 3. We created dataset 3 using dataset 1. We used the control group of dataset 1 as the control group of dataset 3. Then, we changed the expression values of some of the randomly selected genes to model the primary effect of external perturbation. From that perturbed dataset, we simulated the secondary effects using the sigmoid method described in Garg et al. [22]. We denote the parameter for primary perturbation effect by *deviation*. Deviation is the ratio of the change of expression value Δx of a gene to its original expression value x (i.e. $derivation = \frac{\Delta x}{x}$) which is normalized between zero and one. We tuned the other parameters of the method to create a meaningful dataset as follows; $alpha$ = 1, β = 0.01, k_{ac} = 1.0, k_{in} = 1, h = 0.1.

- Dataset 4. We create this dataset from dataset 1 in two steps as follows.

  - Selection of genes. In order to carry out experiments on larger scale data with known PDR genes, we generated data in the presence of a hypothetical perturbation from the real datasets as follows. We first select three sets of genes. Each set consists of some primarily affected genes and a higher number of secondarily affected genes. Here, we describe how we construct each of the three sets of affected genes. We first select a random gene from the network and label it as a primarily affected DE gene. We then traverse its outgoing neighbors in a breadth first search manner. As we visit a gene during traversal, we label it as a secondarily affected DE gene with a probability of $1 - (1 - q)^{\eta}$, where η is the number of incoming DE neighbors. Here q is the probability that a gene is DE due to a DE predecessor (0.4 in our experiments). We repeat these steps to create the desired number of primarily affected genes.

  After we obtain the three set of genes, we assign one set to both D_A and D_B groups. We assign the other two sets of genes to different groups. These two set of genes are differentially regulated as they are affected in only one group and not in the other. The three groups can contain different number of primarily and secondarily affected genes. We call these three sets of genes as primarily differentially regulated, secondarily differentially regulated and equally regulated genes.

  - Generation of gene expression. Once we identify these three sets of genes in the two groups, we create control and non-control data for D_A and D_B over N samples. We use the control part of the real dataset in Smirnov et al. as the control part of our synthetic dataset in both D_A and D_B [20]. To generate the non-control dataset, we traverse each of the genes that participate in the gene networks. Consider a gene g_i with mean and standard deviation of expression in the control dataset 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 but different mean as that of the control group. For the primarily and secondarily affected genes we use $\mu_i \pm d_p$ and $\mu_i \pm d_s$ respectively, where $d_p > d_s$.

  To summarize, we used the same variance in the non-control group as that in the control group. However, for an affected gene we changed the value of the mean in the non-control group from that in the control group. For a primarily affected gene we applied a higher deviation of mean than that of the secondarily affected genes.

Regulatory networks

We collected 24,663 genetic interactions from the 105 regulatory and signaling pathways of KEGG database [23]. Overall 2,335 genes belong to at least one pathway in KEGG. In our model, we considered only the genes that take part in the gene networks.

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 this step 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. We compared the accuracy of CMRF to three such methods namely, SMRF, Student's t test and SSEM.

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 PDR 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 ranks, we created a unified rank of differentially regulated genes. We shall elaborate on this unified rank creation later. We, first, explain how we create ranks on individual groups D_A and D_B for other three methods.

- SMRF. We apply the SMRF to each group separately and obtain a set of differentially expressed genes. We sort the genes in decreasing order of joint likelihood with the metagene. A higher joint likelihood 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 of each group, where it produces a rank for each single data point.
- Student's t test. We use 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. We assume that the null hypothesis corresponds that the gene is EE. So a substantially lower p-value implies a higher chance of being primarily affected. We perform the test on all the genes and rank them according the increasing order of p-values.

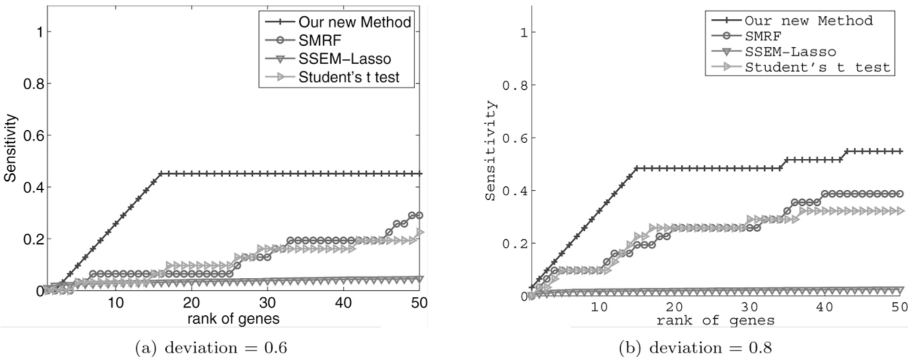
Now we describe how we create an unified ranking of differentially regulated genes for these three methods. We denote the ranks from data group D_A and D_B by R_A and R_B respectively. The unified rank is defined by R_U . We denote the number of genes in each rank to be ω_A and ω_B respectively. We scan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

Results

In this experiment we used dataset 3, that we have just described. To observe the accuracy of CMRF at varying degree of difficulties, we conducted experiments with four different values of *deviation*, namely, {0.5, 0.6, 0.7, 0.8}. However, we discuss only two of them in this paper (see Figure 3) since for other two parameters the results are similar. 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 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 CMRF 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, its achieves a sensitivity of 0.6 when the deviation is 0.8. We obtained similar results for other deviations, which we do not discuss here. The method in SMRF reaches to 30% and 40% accuracy, however at a slower pace. The t-test reaches around 25% and 30% sensitivity at ranking position 50 for these two cases respectively. SSEM's sensitivity is below 0.1 for all experiments even within the top 50 positions.

We believe that there are three major factors for the success of our method over the other competing methods. First, the other methods do not simultaneously handle two groups of datasets and are not able to generate an unified ranking of differentially regulated genes. CMRF encompasses both groups in a single model and probabilistically determines the PDR genes. Hence, it is more shielded against the false positives introduced during the unification of ranking. Second, CMRF can successfully incorporate the gene interactions using MRFs while others ignore this information. Finally, in real perturbation experiments, multiple genes are often primarily affected. CMRF is capable of dealing with both large and small number of primarily affected genes, while performances of other methods deteriorate as the number of primarily affected genes grows. Thus, we conclude that our method is more suitable for real perturbation experiments.



(a) deviation = 0.6

(b) deviation = 0.8

Figure 3 Comparison of CMRF with three other methods. Comparison of our method CMRF to SMRF, SSEM and t-test. The number of primarily differentially affected genes is 40. The values for deviation (maximum perturbation to the PDR genes) are 0.6 and 0.8. The figures indicate that CMRF outperforms SMRF, SSEM and t-test.

Statistical significance experiment

The experiments in the last section enable us to compare the accuracy of CMRF with that of the other methods on synthetic datasets. We also wanted to evaluate the accuracy of CMRF on real dataset. However, we do not have any gold standard available that enlists *true* set of PDR genes. Hence, we conducted a set of statistical significance experiments to estimate the confidence of our accuracy. Specifically, we obtained the control data from a real dataset, perturbed it in a controlled way for a number of genes. We calculated the likelihood probabilities of those genes and created a distribution. We repeated this process with varying amount of perturbation. Finally, we executed CMRF on a real dataset and analyzed the result.

Results

We obtained the real dataset from drug response experiment conducted by Taylor et al [3], which is actually dataset 2. Apart from this real dataset, we create different versions of dataset 4 by varying d_p as {0.1, 0.2, 0.3,..., 3.0}. If $d_p > 1.1$, we set d_s to 1, otherwise $d_s = 0.5 \times d_p$. Thus, we have 30 synthetic datasets in total. In every dataset, we fix the number of primarily and secondarily differentially regulated genes to 50 and 172 respectively. To decide whether a gene g_i is DR, when g_i is DE in D_A and EE in D_B, we define a null-hypothesis H_{0i}: g_i is *DR, but in the reverse way, i.e.* g_i *is EE in* D_A *and DE in* D_B. We calculate the likelihood of being EE in D_A and DE in D_B for that gene, as described. For gene g_i, we denote the log likelihood of accepting H_{0i} by LL_i. In every dataset, we create a box plot of the 50 LL_i values, as the number of DR genes in each dataset is 50. A lower value LL_i indicates that g_i has a higher probability of being differentially regulated.

Figure 4 illustrates the statistical significance of the experiments over the datasets with $d_p = 1.2$ to 2.0. The box plot demonstrates a relationship between the P-value and d_p. A higher value of d_p indicates a lower P-value and hence, a high chance of being PDR. We also observe that the variance of P-value increases with the increase of d_p.

We also executed CMRF on the real datasets without any modification. Interestingly, on the real dataset from Taylor et al. [3] (dataset 2), we did not obtain any genes as differentially regulated. A careful observation concludes that when both the number of data points and the gap d_p (i.e. the signal to noise ratio) is low, the coefficients γ_6 and γ_7 in the prior density become strong and all genes are identified as equally regulated. However, when either the number of data points or d_p is significantly high, the data can overcome the prior. In the current real dataset, the number of data points is only 33 and the gaps between the control and non-control group were less than $1.2 \times \sigma$. As a result, CMRF

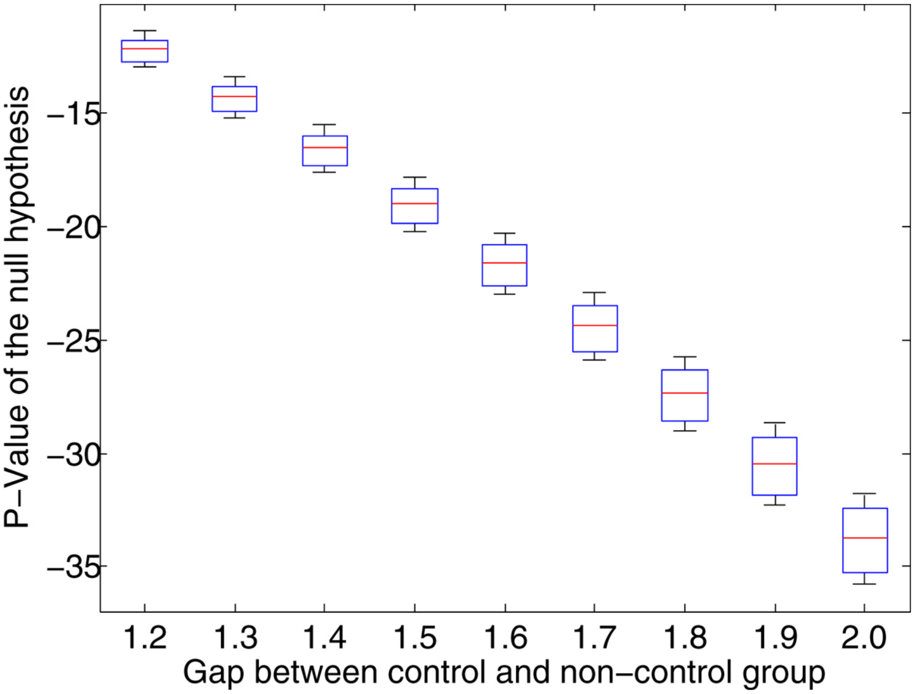


Figure 4 Illustration of statistical significance test. Illustration of the statistical significance test. Box plot demonstrates the P-value of null hypothesis of the DR genes over synthetic dataset. From the plot we clearly conclude that a higher gap between the control and non-control group of a DR gene leads to a lower P-value. The genes with a lower P-value have a higher chance of being primarily differentially regulated.

identifies no differentially regulated genes in the dataset. Thus, we can conclude that either there is not much difference between the two groups in the real data, or the data does not contain enough data points, so that our model can highlight difference between the two groups.

In Figure 4 we present the results for d_p = 1.2 to 2.0. Note, that for $d_p < 1.2 \times \sigma$ our model did not identify any DR genes. Here also, we attribute a similar reason for not finding any DR genes as both d_p and the number of data points are small. On the other hand, in Section Comparison to other methods, when we execute CMRF on synthetic datasets with 155 data points we were able to identify a substantial number of true PDR genes even with $d_p = 0.02 \times \sigma$. To substantiate our conclusion that there exists little difference between the two groups in the real dataset, we conducted a set of permutation tests. We shuffled the two original groups to create new sets of data. We repeated this process for a number of times (40 in the present experiment) and executed CMRF on each of them. For every derived dataset, CMRF did not find any DR genes. Hence, this experiment bolsters the claim that there are no DR genes in the original real data.

An interesting question can be raised is that "is there indeed no DR genes in the real dataset from Taylor et al. [3]?" Another similar question can be "will our method be able to detect DR and PDR genes from similar other real datasets?" We believe that CMRF requires a bigger dataset for DR and PDR genes to be discovered. For example, CMRF is able to identify the DR and PDR genes from the synthetic dataset that contains substantially higher number of data points than that of the real dataset. Since the difference between control and noncontrol groups of a DE gene is small compared to the variance of the data points, it is difficult to detect that subtle effect of perturbation with a small dataset. For a small dataset, the prior due to third hypothesis becomes strong and the two corresponding parameters γ_6 and γ_7 assumes extreme values. Thus the support from data is not sufficient to overcome the prior and hence, the method is not able to identify the DR and PDR genes. There are two solutions to overcome this problem. First of them is to employ a bigger dataset. With the advancement of comparatively inexpensive and high throughput technologies bigger dataset are increasingly common nowadays. From that perspective, CMRF is supposed to perform more accurately in the near future. A second option to circumvent the problem is to restrict the growth of the two parameters γ_6 and γ_7. If we have knowledge about the values of these two parameters, we can assign then as input to the program and refrain from estimating their values. This will enable us to employ a comparatively non-informative prior which will be easier for the data to overcome. Also, we can use specific bound over those variables while estimating them to avoid them becoming stronger.

Conclusions

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 solved the problem of finding primarily differentially regulated genes in the presence of external perturbations when the data is sampled from two groups. The probabilistic Bayesian method based on Markov Random Field incorporates dependency structure of the gene networks as the prior to the model. Experimental results on synthetic and real datasets demonstrated the superiority of CMRF compared to other simple techniques.

Methods

In this section we describe different components of CMRF. Section Notation and problem formulation describes the notation and formulates the problem. Section Overview of the solution provides a high level overview of the solution. Section Computation of the prior density function describes the calculation of the prior density function of MRF. Section Approximation of the objective function discusses the definition of a tractable objective function. Section Computation of likelihood density function discusses the calculation of the likelihood function. Finally, Section Objective function optimization describes the algorithm to optimize the objective function.

Notation and problem formulation

In this section, we describe our notation and formally define the problem. We define a Bayesian model for gene expression in a two-group perturbation experiment. We classify the random variables of the model into two different groups, namely *observed variables* and *hidden variables*. We have the values for the observed variables, while we estimate the values of the 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 $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$. For each gene and for each group the microarray data contains the gene expression values before and after the perturbation, i.e. control and non-control data respectively. We denote the expression value of the ith gene from the jth sample in the control data of group D_A with y_{Aij}. We represent the same for the non-control data with y'_{Aij}. Thus the expression values of the gene g_i for all the samples in D_A for control and non-control data are $\mathbf{y_{Ai}} = \{y_{Ai1}, y_{Ai2}, \cdots y_{AiN_A}\}$ and $\mathbf{y_{Ai}}' = \{y_{Ai1}', y_{Ai2}', \cdots y_{AiN_A}'\}$ respectively. We denote all the expression values in group D_A for gene g_i with Y_{Ai}(i.e. $Y_{Ai} = \mathbf{y_{Ai}} \cup \mathbf{y'_{Ai}}$). We denote the collection of the gene expressions of all the genes in group D_A by $\mathcal{Y}_A = \bigcup_{i=1}^{M} Y_{Ai}$. We define $\mathcal{Y}_B$ similarly for all the genes in D_B. We refer the complete gene expression data using variable $\mathcal{Y} = \mathcal{Y}_A \cup \mathcal{Y}_B$.

- Neighborhood variables. We use the term $\mathcal{W} = \{W_{ij}\}$ to indicate if 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 1 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, $X_{ij} = 4 \times (Z_i - 1) + Z_j$. Note that, this equation is created to maintain brevity of the mapping between the interaction variables and the regulation variables by carefully assigning different numeric constants between one and 16 to appropriate values of an interaction variable. Table 1 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.

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} is based on the values of four independent variables S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj} . Table 1 enumerates the values of Z_i, Z_j and X_{ij} for different values of S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj} .

It is worth noting that the different values that we assign to the hidden variables are categorical in nature.

Problem formulation

Let $\mathcal{G} = \{g_1, g_2, \cdots, g_M\}$ denote the set of all genes. Using the definition 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 $(X_{0j}$ $|\cdot)$, $\forall_j \in$ $\{1, 2,... M\}$, we can create an ordered list of candidate PDR genes.

Overview of the solution

This section describes a high level overview of our approach to estimate $p(X_{0j}|\cdot)$, $\forall_j \in$ $\{1, 2,... M\}$. One simple approach can be using a hypothesis test to find out the PDR genes in the given dataset [15]. However, the available hypothesis tests do not consider the interactions among genes in the gene network. Also, deciding on the significance of test can be a complex step. Another approach can be to use SSEM to create a rank

Table 1 Enumeration of the values of Z_i, Z_j and X_{ij}

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

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 turn depends on the values of four independent variables S_{Ai}, S_{Bi}, S_{Aj} and S_{Bj}.

of the potential primarily affected genes in each group separately [11]. Then we can select the top k genes in each group and perform a set difference to obtain the PDR genes. Though SSEM considers the correlation between the genes, it does not utilize any known information from the gene networks.

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 [24] we can write the joint probability density of interaction variables $\mathcal{X}$ as,

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

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. A detailed discussion of how we compute this likelihood can be found in Section Computation of likelihood density 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 Computation of the prior density function. 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.

A Maximum Likelihood Estimation (MLE) of Equation 1 is practically infeasible even for a small number of genes since the number of terms in the denominator grows exponentially. Instead we use a pseudo-likelihood version of the objective function as shown in Section Approximation of the objective function. We use Iterative Conditional Modes (ICM) [19] and Differential Evolution [25] 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 $p(X_{0j}|\cdot)$, $\forall_j \in \{1, 2, \dots M\}$, we create an ordered list of candidate PDR genes. We elaborate on each of these steps next.

Figure 2 illustrates different portions of CMRF and the connectivity between them.

Computation of the prior density function

In this section, we describe how 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. We present the entire concept in three numbered subsections.

1. Statement of hypotheses

Here we state the three hypotheses on the biological networks in brief.

- Hypothesis 1. In each group $D_T(T \in \{A, B\})$, the metagene g_0 can change the state of all the other genes. Thus, all the genes can be directly affected by the external perturbation.
- Hypothesis 2. 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, i.e. a gene can be indirectly affected by the perturbation through genetic interactions.
- Hypothesis 3. 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. We expect that the response of genes in these groups is very similar.

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.

2. Markov Random Field construction

In order to compute the prior density function, we define a Markov Random Field (MRF) over the $\mathcal{X}$ variables [18]. 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 $\Psi = (\mathcal{X},\ \mathcal{E})$, 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 5(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 5(a).

One important point to note is that, this graph does not use the state variables $\mathcal{S}$ or the regulation variables $\mathcal{Z}$ to model the dependencies between the genes. Rather, it establishes those dependencies over the $\mathcal{X}$ variables. For example, in Figure 5(b) we draw the MRF graph corresponding to the hypothetical gene network in Figure 5(a). In the gene network, there is an edge

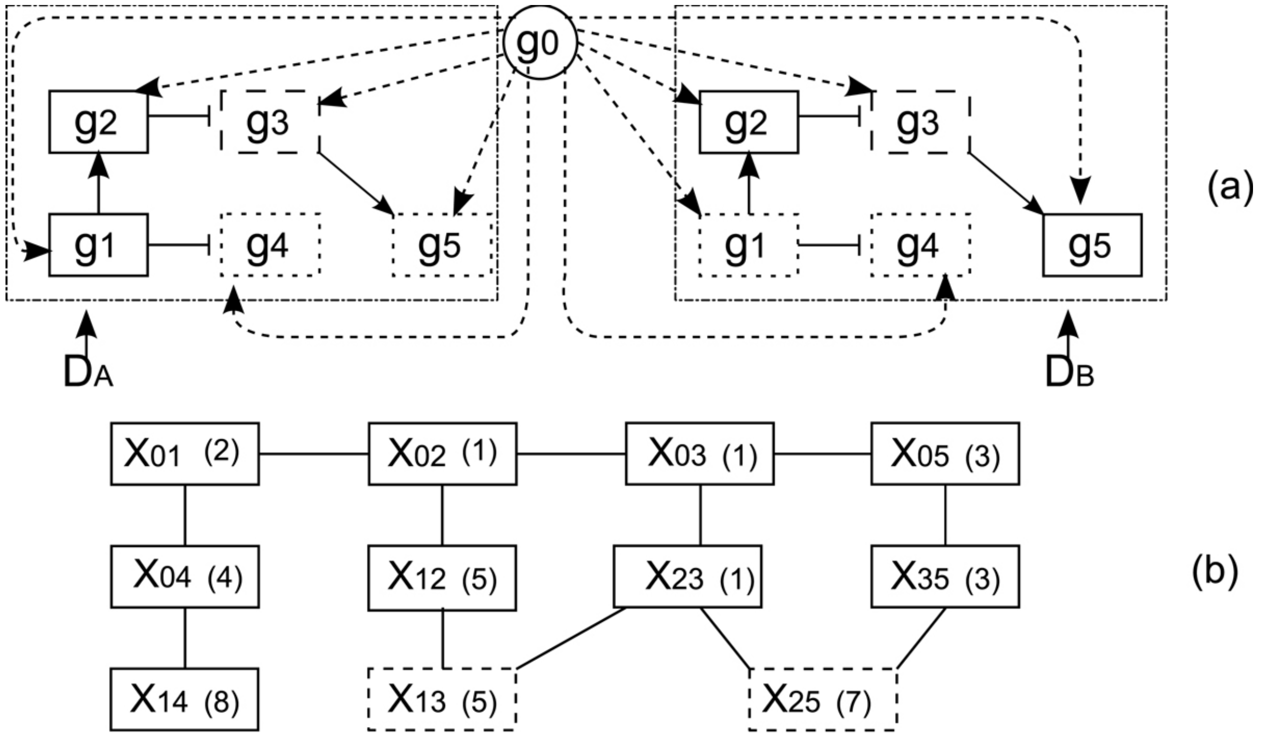


Figure 5 A hypothetical gene network and corresponding Markov random graph. (a) A small hypothetical gene network with perturbation in two datasets D_A and D_B. The genes in the two datasets interact through identical network, although they assume different states. The circle g_0 represents the abstraction of the external perturbation. Rectangles denote genes. → implies activation and ⊣ 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 Field 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.

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 g1 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 same state (i.e. $S_{T2} = S_{T3}, T \in \{A, B\}$). We formulate these dependency constraints using a set of unary and binary functions called *feature functions*. We discuss these feature functions next.

3. Development of feature functions

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 clique C_{ij}. This function evaluates to one or zero, 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 defined over that MRF as, $p(\mathcal{X}|\theta_X) = \frac{1}{\Delta}\prod_{C_{ij},W_{ij}=1}\psi(C_{ij})$ [26]. 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 hypotheses.

Unary feature functions

F$_1$, F2, F3. A primary component of the prior density function is modeling the frequency of X_{ij} itself. Here, we focus 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 $X_{ij} = 2$. It returns 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}

Table 2 Feature functions

X_{ij}	F_1	F_2	F_3	F_6	F_7
1	0	0	1	1	1
2	1	0	0	1	0
3	0	1	0	1	0
4	0	0	1	1	1
5	0	0	1	0	1
6	0	0	1	0	0
7	0	0	1	0	0
8	0	0	1	0	1
9	0	0	1	0	1
10	0	0	1	0	0
11	0	0	1	0	0
12	0	0	1	0	1
13	0	0	1	1	1
14	0	0	1	1	0
15	0	0	1	1	0
16	0	0	1	1	1

Enumeration of five different unary feature functions F_1, F_2, F_3, F_6 and F_7.

by a feature function called $F_3(X_{ij})$. It returns zero when $X_{ij} \in \{2, 3\}$ and equals to one otherwise. Table 2 enumerates the the domains and ranges of F_1, F_2 and F_3.

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°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 [3].

Table 3 Left external equality

									X_{pj}								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
	1	1				0				0				0			
	2		1				0				0				0		
	3			1				0				0				0	
	4				1				0				0				0
	5	0				1				0				0			
	6		0				1				0				0		
	7			0				1				0				0	
X_{ij}	8				0				1				0				0
	9	0				0				1				0			
	10		0				0				1				0		
	11			0				0				1				0	
	12				0				0				1				0
	13	0				0				0				1			
	14		0				0				0				1		
	15			0				0				0				1	
	16				0				0				0				1

The table enumerates the truth values for the binary feature function *left external equality* (f_4). Only the possible entries are annotated with zero and one. The other entries require different values of Z_j in X_{ij} and Xp_j, which is not possible. Note, that the feature function can assume one only when X_{ij} and X_{pj} are equal, which is in accordance with the definition of that feature function.

Table 4 Right external equality

									X_{ik}								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
	1	1	0	0	0												
	2	0	1	0	0												
	3	0	0	1	0												
	4	0	0	0	1												
	5					1	0	0	0								
	6					0	1	0	0								
	7					0	0	1	0								
X_{ij}	8					0	0	0	1								
	9									1	0	0	0				
	10									0	1	0	0				
	11									0	0	1	0				
	12									0	0	0	1				
	13													1	0	0	0
	14													0	1	0	0
	15													0	0	1	0
	16													0	0	0	1

The table enumerates the truth values for the binary feature function *right external equality* (f_5). Only the possible entries are annotated with zero and one. The other entries require different values of Z_i in X_{ij} and X_{ik}, which is not possible. Note, that the feature function can assume the value one only when X_{ij} and X_{ik} are equal, which is in accordance with the definition of right external equality.

$\Upsilon^{\circ}$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 → 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°holds true. Similarly, for the second set of edges X_{ij} equals to X_{ik} *iff* $Z_j = Z_k$, which in tern implies that $\Upsilon^{\circ}$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,

$$F_5(X_{ij}) = \sum_{k,\, W_{ij}=1, W_{jk}=1} f_5(X_{ij},\ X_{ik}).$$

Tables 3 and 4 enumerate the values of f_4 and f_5 for different values of X_{ij}. The missing entries in these tables correspond to the cases which can not occur during the optimization. For instance, in Table 3, a missing entry corresponds to different values of Z_j in X_{ij} and X_{pj} which is not possible.

For feature functions f_4 and f_5, 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 dotted rectangles in Figure 5(b).

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,

$$F_6(X_{ij}) = \sum_{i,j,W_{ij}=1,t\in\{1,\cdot,4,\ 13,\cdots,16\}} f_6(X_{ij},\ t).$$

- 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,

$$F_7(X_{ij}) = \sum_{i,j,W_{ij}=1,t\in\{1,4,5,8,9,12,13,16\}} f_7(X_{ij},\ t).$$

The last two columns of Table 2 enumerate these two internal equalities.

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

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

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

In the next section, we discuss how we approximate the objective function of the MRF and the data. We also describe how we formulate the posterior probability density function for X_{ij}.

Approximation of the objective function

A direct maximization of the objective function given by Equation 1 is intractable, as it requires evaluation of exponential number of terms in the denominator. We employ pseudo-likelihood as an established substitute to Equation 1 [27]. 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 [28]. The approximated objective function can be written as,

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

The posterior density function F_{ij} of X_{ij} as,

$$\begin{aligned} F_{ij} &= p(X_{ij}|\mathcal{X} - X_{ij},\ \mathcal{Y},\ \theta_X,\ \theta_Y) \\ &= \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}|X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}|\mathcal{X} - X_{ij}, \theta_X)}{\sum_{X_{ij}\in\{1,\cdots,16\}} p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}|X_{ij}, X^*_{ij}, \theta_Y)} \end{aligned} \qquad (4)$$

Derivation of $\boldsymbol{F_{ij}}$.

$$\begin{aligned}
& F_{ij} \\
& = p(X_{ij}\,|\,\mathcal{X} - X_{ij},\ \mathcal{Y},\ \theta_X,\ \theta_Y) \\
& = p(X_{ij}\,|\,\mathcal{X} - X_{ij},\ Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj},\ \theta_X,\ \theta_Y) \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \mathcal{X} - X_{ij} - X^*_{ij}, X_{ij}, X^*_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \mathcal{X} - X_{ij} - X^*_{ij}, X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \mathcal{X} - X_{ij} - X^*_{ij}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \mathcal{X} - X_{ij} - X^*_{ij}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij} - X^*_{ij}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij} - X^*_{ij}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij} - X^*_{ij}, X_{ij}, X^*_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij} - X^*_{ij}, X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_X, \theta_Y)p(\mathcal{X} - X_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \theta_X\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}, \theta_X\,|\,X^*_{ij}, \theta_Y)p(X^*_{ij}, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(\theta_X\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(\theta_X\,|\,X^*_{ij}, \theta_Y)p(X^*_{ij}, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_Y)p(X_{ij}, X^*_{ij}, \theta_X, \theta_Y)p(X^*_{ij}, \theta_X, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X, \theta_Y)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X)}{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X^*_{ij}, \theta_Y)} \\
& = \frac{p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)p(X_{ij}\,|\,\mathcal{X} - X_{ij}, \theta_X)}{\sum_{X_{ij}\in\{12,3,\cdots16\}} p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj}\,|\,X_{ij}, X^*_{ij}, \theta_Y)}
\end{aligned}$$

In step 2 of the derivation, we substitute $\mathcal{Y}$ by Y_{Ai}, Y_{Bi}, Y_{Aj} and Y_{Bj} as X_{ij} is independent of all Y_{Ck} such that $k \neq \{i, j\}$ and $C \neq \{A, B\}$. Also, in the $15th$ step we assume that X_{ij} is independent of θ_Y given $\mathcal{X} - X_{ij}$ and θ_X. □

Derivation of $p(X_{ij}|\mathcal{X} - X_{ij}, \theta_X)$, $W_{ij} = 1$.

$$\begin{aligned} & p\,(X_{ij}|\mathcal{X} - X_{ij},\ \theta_X) \\ & = \frac{p(\mathcal{X}, \theta_X)}{P(\mathcal{X} - X_{ij}, \theta_X)} \\ & = \frac{p(\mathcal{X}, \theta_X)}{\sum_{X_{ij} \in \{1,2,3\ \cdots\ 16\}} P(\mathcal{X} - X_{ij}, X_{ij}, \theta_X)} \\ & = \frac{A(X_{ij}).B(ij)}{\sum_{t=\{1,2,3,\ \cdots\ 16\}} A(t) \cdot B(ij)} \end{aligned}$$

$A\ (X_{ij})$ is $exp(\sum_{k \in \{1,2,\cdots,7\}} \gamma_k F_k(X_{ij}))$ and $B(ij)$ is given by $exp(\sum_{m,n,ij \neq mn,k \in \{1,2,\cdots,7\}} \gamma_k F_k(X_{mn}))$. Here, we denote the prior density parameters $\{\gamma_1, \gamma_2, \ldots \cdot \gamma_7\}$ by θ_X. Canceling $B(ij)$ from numerator and denominator the density function simplifies to,

$$\begin{aligned} & p(X_{ij}\,|\,\mathcal{X} - X_{ij},\ \theta_X) \\ & = \frac{exp(\sum_{k \in \{1,2,\cdots,7\}} \gamma_k F_k(X_{ij}))}{\sum_{t=\{1.2.3,\cdots 16\}} exp(\sum_{k \in \{1,2,\cdots,7\}} \gamma_k F_k(X_{ij} = t))} \end{aligned}$$

□

There are two different terms in objective function of Equation 4. $p(X_{ij}|\mathcal{X} - X_{ij}, \theta_X)$ stands for the conditional prior density function of X_{ij} which we just have derived from using Bayes rule. In the next section, we discuss the likelihood function $p(Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj}|X_{ij},\ X^*_{ij},\ \theta_Y)$.

Computation of likelihood density function

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

1. Likelihood for a single gene

Consider a set of measurements for a gene g_i that follows a single Gaussian distribution by $\mathbf{z_i} = \{z_{i1}, z_{i2}, \ldots, z_{iN}\}$. We denote the latent mean of $\mathbf{z_i}$ as μ and the standard deviation as σ. As different genes can have different average expressions, we assume that μ follows a genome wise distribution with mean μ_0 and standard deviation τ [29]. Thus, for $\mathbf{z_i}$, the likelihood for the data points in that group is given by,

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

The derivation of Equation 5 can be obtained from Demichelis et al [30]. If a gene is DE, its expression measurements in control and non-control groups follow different distributions [29]. On the other hand, for equally expressed genes, all the measurements in both groups share the same mean. The likelihood function for a DE gene g_i in group D_T ,$T \in \{A, B\}$ is given by,

$$\mathcal{L}_{\mathrm{T_{DE}}}(g_i) = L(\mathbf{y_i}\,|\,\mu_0,\ \sigma^2,\ \tau^2)L(\mathbf{y'_i}\,|\,\mu_0,\ \sigma^2,\ \tau^2) \qquad (6)$$

Similarly, for EE genes it is given by,

$$\mathcal{L}_{\mathrm{T_{EE}}}(g_i) = L(\mathbf{y_i} \cup \mathbf{y'_i}\,|\,\mu_0,\ \sigma^2,\ \tau^2) \qquad (7)$$

For instance, the likelihood of a gene to be DE in group D_A is given by $\mathcal{L}_{\mathrm{A}_{DE}}(g_i)$.

2. Likelihood for a regulation variable

As for a gene g_i, the regulation variable Z_i can assume four different values from 1 to 4, the equations of the likelihood that a gene is DR or ER also take four different forms given by,

$$\mathcal{L}_Z(g_i) = \begin{cases} \mathcal{L}_{\mathrm{A_{DE}}}(g_i)\ \mathcal{L}_{\mathrm{B_{DE}}}(g_i), & \text{if } Z_i = 1. \\ \mathcal{L}_{\mathrm{A_{DE}}}(g_i)\ \mathcal{L}_{\mathrm{B_{EE}}}(g_i), & \text{if } Z_i = 2 \\ \mathcal{L}_{\mathrm{A_{EE}}}(g_i)\ \mathcal{L}_{\mathrm{B_{DE}}}(g_i), & \text{if } Z_i = 3 \\ \mathcal{L}_{\mathrm{A_{EE}}}(g_i)\ \mathcal{L}_{\mathrm{B_{EE}}}(g_i), & \text{if } Z_i = 4 \end{cases}$$

2. Likelihood for an interaction variable

We have 16 different forms for the likelihood of the X_{ij} due to its 16 different values. However, here, we shall derive only for $X_{ij} = 1$, as for the other values of X_{ij} we have a similar derivation.

$$\begin{aligned} & p(Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj}\,|X_{ij} = 1,\ X^*_{ij},\ \theta_Y) \\ & = \sum_{\tau_i,\tau_j \in \{1,\cdots,4\}} p(Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj}\,|Z_i = \tau_i,\ Z_j = \tau_j,\ \theta_Y). \\ & \quad p(Z_i = \tau_i,\ Z_j = \tau_i,\ \theta_Y\,|X_{ij} = 1,\ X^*_{ij},\ \theta_Y) \end{aligned} \qquad (8)$$

From the definition of X_{ij}, $p(Z_i = \tau_i,\ Z_j = \tau_i,\ \theta_Y\,|X_{ij} = 1,\ X^*_{ij},\ \theta_Y)$ equals to 1 when $Z_i = 1$ and $Z_j = 1$. Its value is zero for all other values of Z_i and Z_j. So, continuing from the last step of Equation 8,

$$\begin{aligned} & p(Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj}\,|X_{ij} = 1,\ X^*_{ij},\ \theta_Y) \\ & = p(Y_{Ai},\ Y_{Bi},\ Y_{Aj},\ Y_{Bj}\,|Z_i = 1,\ Z_j = 1,\ \theta_Y) \\ & = p(Y_{Ai},\ Y_{Bi}\,|Z_i = 1,\ Z_j = 1,\ \theta_Y). \\ & \quad p(Y_{Aj},\ Y_{Bj}\,|Z_i = 1,\ Z_j = 1,\ \theta_Y) \\ & = p(Y_{Ai},\ Y_{Bi}\,|Z_i = 1,\ \theta_Y)p(Y_{Aj},\ Y_{Bj}|Z_j = 1,\ \theta_Y) \\ & = \mathcal{L}_Z(g_i)\mathcal{L}_Z(g_j) \end{aligned} \qquad (9)$$

In a similar way, we can derive the likelihood functions for all the 16 different values of X_{ij} variables. A special case arises when g_i is the metagene, i.e. g_0. We assume that $\mathcal{L}_{\mathrm{T}_{\mathrm{DE}}}(g_0) = 1$ and $\mathcal{L}_{\mathrm{T}_{\mathrm{EE}}}(g_0) = 0$, $T \in \{A, B\}$. Thus, the likelihood of the metagene given $Z_0 = 1$ equals to 1. Its value is zero otherwise.

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 $\theta_X = \{\gamma_1,... \gamma_7\}$ 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 $\theta_X = \{\gamma_1,... \gamma_7\}$. We use a global optimization method called differential evolution to optimize both of them [25]. To optimize the objective function in Equation 3, we employ the ICM algorithm described by Besag [19]. 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,

$$\arg\max_{\theta_Y} \prod_{X_{ij}, W_{ij}=1} p(Y_{Ai}, Y_{Bi}, Y_{Aj}, Y_{Bj} | X_{ij}, X^*_{ij}, \theta_Y)$$

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,

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

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 $\prod_{X_{mn}} p(X_{mn}|\mathcal{X} - X_{mn}, \mathcal{Y}, \theta_X, \theta_Y)$ when $X_{mn} \in \{X_{rt} | r = i \text{ or } t = i, W_{rt} = 1\}$.

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. Figure 2 illustrates different components of CMRF and the connectivity between them.

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.

Acknowledgements
This work was supported partially by NSF under grants CCF-0829867 and IIS-0845439.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 2, 2012: Selected articles from the First IEEE International Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S2.

Authors' contributions
NB conceived the study, analyzed the data, implemented the methods, supplied the analysis tools, designed the experiments, performed the experiments and wrote the paper. MS conceived the study and participated in writing the paper. SR conceived the study, designed the experiments and participated in writing the paper. TK conceived the study, designed the experiments and participated in writing the paper.

Competing interests
The authors declare that they have no competing interests.

Published: 12 April 2012

References

1. Cheng R, Zhao A, Alvord W, Powell D, Bare R, Masuda A, Takahashi T, Anderson L, Kasprzak K: **Gene expression dose-response changes in microarrays after exposure of human peripheral lung epithelial cells to nickel(II).** *Toxicol Appl Pharmacol* 2003, **191**:22-39.
2. Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, Eng J, Bumgarner R, Goodlett D, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292(5518)**:929-34.
3. Taylor K, Pena-Hernandez K, Davis J, Arthur G, Duff D, Shi H, Rahmatpanah F, Sjahputera O, Caldwell C: **Large-scale CpG methylation analysis identifies novel candidate genes and reveals methylation hotspots in acute lymphoblastic leukemia.** *Cancer Res* 2007, **67(6)**:2617-25.
4. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-126.
5. Marton MJ, Derisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, Bassett DE, Hartwell LH, Brown PO,

Friend SH: **Drug target validation and identification of secondary drug target effects using DNA microarrays.** *Nat Med* 1998, **4(11)**:1293-1301.
6. Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW: **Genomic profiling of drug sensitivities via induced haploinsufficiency.** *Nature Genetics* 1999, **21(3)**:278-283.
7. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW: **Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101(3)**:793-798.
8. Lum P, Armour C, Stepaniants S, Cavet G, Wolf M, Butler J, Hinshaw J, Garnier P, Prestwich G, Leonardson A, Garrett-Engele P, Rush C, Bard M, Schimmack G, Phillips J, Roberts C, Shoemaker D: **Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes.** *Cell* 2004, **116**:121-37.
9. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C: **Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways.** *Nature Biotechnology* 2003, **22**:62-69.
10. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ: **Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.** *Nature Biotechnology* 2005, **23(3)**:377-383.
11. Cosgrove EJ, Zhou Y, Gardner TS, Kolaczyk ED: **Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia.** *Bioinformatics* 2008, **24(21)**:2482-2490.
12. Bandyopadhyay N, Somaiya M, Kahveci T, Ranka S: **Modeling perturbations using gene networks.** *Proc LSS Comput Syst Bioinform Conf* 2010, 26-37.
13. Conesa A, Nueda MJ, Ferrer A, Talón M: **maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments.** *Bioinformatics* 2006, **22(9)**:1096-1102.
14. Hong F, Li H: **Functional Hierarchical Models for Identifying Genes with Different Time-Course Expression Profiles.** *Biometrics* 2006, **62(2)**:534-544.
15. Chuan Tai Y, Speed TP: **On Gene Ranking Using Replicated Microarray Time Course Data.** *Biometrics* 2009, **65**:40-51.
16. Angelini C, De Canditiis D, Pensky M: **Bayesian models for two-sample time-course microarray experiments.** *Computational Statistics & Data Analysis* 2009, **53(5)**:1547-1565.
17. Van Deun K, Hoijtink H, Thorrez L, Van Lommel L, Schuit F, Van Mechelen I: **Testing the hypothesis of tissue selectivity: the intersection-union test and a Bayesian approach.** *Bioinformatics* 2009, **25(19)**:2588-2594.
18. Li SZ: *Markov Random Field Modeling in Image Analysis.* 3 edition. Springer Publishing Company, Incorporated; 2009.
19. Besag J: **On the statistical analysis of dirty pictures.** *Journal of the Royal Statistical Society* 1986, **48(3)**:259-302.
20. Smirnov D, Morley M, Shin E, Spielman R, Cheung V: **Genetic analysis of radiation-induced changes in human gene expression.** *Nature* 2009, **459(7246)**:587-91.
21. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J, White R: **Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome.** *Genomics* 1990, **6(3)**:575-7.
22. Garg A, Mendoza L, Xenarios I, DeMicheli G: **Modeling of multiple valued gene regulatory networks.** *Conf Proc IEEE Eng Med Biol Soc* 2007, **2007**:1398-404.
23. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
24. Bishop CM: *Pattern Recognition and Machine Learning (Information Science and Statistics)* Secaucus, NJ, USA: Springer-Verlag New York, Inc; 2006.
25. Storn R, Price K: **Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces.** *Journal of Global Optimization* 1997, **11(4)**:341-359.
26. Hammersley JM, Clifford P: **Markov fields on finite graphs and lattices.** *Unpublished manuscript* 1968.
27. Besag J: **Efficiency of pseudolikelihood estimation for simple Gaussian fields.** *Biometrika* 1977, **64(3)**:616-618.
28. Geman S, Graffigne C: **Markov random field image models and their applications to computer vision.** *Proceedings of the International Congress of Mathematics: Berkley* 1986, 1496-1517.
29. Kendziorski C, Newton M, Lan H, Gould M: **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.** *Stat Med* 2003, **22(24)**:3899-914.
30. Demichelis F, Magni P, Piergiorgi P, Rubin M, Bellazzi R: **A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays.** *BMC Bioinformatics* 2006, **7**:514.

doi:10.1186/1471-2164-13-S2-S2
Cite this article as: Bandyopadhyay *et al.*: **CMRF: analyzing differential gene regulation in two group perturbation experiments.** *BMC Genomics* 2012 **13**(Suppl 2):S2.