Statistica Sinica **18**(2008), 709-729

A NONPARAMETRIC EMPIRICAL BAYES APPROACH TO JOINT MODELING OF MULTIPLE SOURCES OF GENOMIC DATA

Wei Pan[1], Kyeong S. Jeong[2], Yang Xie[3] and Arkady Khodursky[1]

[1] *University of Minnesota,* [2] *University of California, Los Angeles*
and [3] *University of Texas Southwestern Medical Center*

Abstract: With the rapid accumulation of various high-throughput genomic and proteomic data, one is compelled to develop new statistical methods that can take advantage of existing multiple sources of data. In our motivating example, a chromatin-immunoprecipitation (ChIP) microarray experiment was conducted to detect binding target genes of a broad transcription regulator, leucine responsive regulatory protein (Lrp) in *E. coli.* In addition, a cDNA microarray dataset is available to compare gene expression of the wild type with that of a mutant with the Lrp gene deleted in *E. coli.* It is biologically reasonable to assume that the genes with altered expression are more likely to be regulated by Lrp than those with no expression change. Hence we aim to borrow information in the gene expression data to increase statistical power to detect the binding targets of Lrp. We propose a novel joint model for protein-DNA binding data and gene expression data; under mild modeling assumptions, it is shown that the method is optimal, equivalent to a joint likelihood ratio test. We compare the joint modeling with two existing methods of combining separate analyses. We adopt a nonparametric empirical Bayes (EB) method to draw statistical inference in the joint model; in particular, we propose a new method, maximum likelihood conditional on the binding data, to estimate two prior probabilities for the expression data, which are non-identifiable based on the expression data alone. We use simulated data to demonstrate the improved performance of the joint modeling over other approaches. Application to the Lrp data also shows better performance of the joint modeling than that of analyzing the binding data alone.

Key words and phrases: ChIP-chip, computational biology, false discovery rate, gene expression, Lrp, microarray.

1. Introduction

High-throughput biotechnologies, such as microarrays, have generated large amounts and various types of genomic and proteomic data. A widespread use of microarray experiments is to monitor genome-wide gene expression. Gene expression or transcription is the process of genetic information flow from DNA sequence to messenger RNA (mRNA). Although any cell of an organism contains all the necessary DNA information for gene expression, obviously, at any given moment, not all the genes in the cell are equally expressed, or even expressed at all: some genes are switched on while others are off, and those switched on may have different expression levels. Gene expression largely depends on physiological and environmental conditions of the cell at the moment. Biologically, a fundamental question is how gene expression is regulated. A general mechanism is through some regulatory proteins called transcription factors (TFs): a TF binds to one or more specific DNA subsequences in a gene's promoter region, called binding sites or motifs, then works with other TFs to stimulate or inhibit the expression of the gene. Although significant progress has been made in understanding a few specific examples, as well as the general aspect of gene transcription regulation, it remains largely unknown which TFs regulate which genes. The biological goal of this work is to discover the target genes of a TF, not necessarily its binding sites or motifs. The most popular approach to identifying bindings sites of a TF is by computationally predicting motifs (i.e., specific DNA sequences that a TF binds to) through DNA sequence alignment (e.g., Liu, Neuwald and Lawrence (1999)). However, presence of a motif in a gene's control region does not necessarily imply that the TF indeed binds to the site *in vivo*. A new application of microarray technology is to identify *in vivo* genome-wide binding locations of a TF via chromatin-immunoprecipitation (ChIP) (e.g., Ren et al. (2000)). Because the resulting DNA-protein binding data are in the usual format of cDNA microarray gene expression data, it is technically possible to apply any of many existing statistical methods of detecting differential gene expression to binding data (Lee et al. (2002)), see Pan (2002) and Smyth, Yang and Speed (2003) for reviews on statistical analysis of gene expression data. Due to high noise in microarray data and typically few replicates, any statistical method being applied may yield results with relatively high false positives or high false negatives. To maximize statistical power, we can take advantage of the existence of other sources of data, and thus their contained information. One source is gene expression data. In our motivating example, on the one hand we have a DNA-protein binding data set to detect the binding targets of a broad transcription regulator, leucine responsive regulatory protein (Lrp), on the other hand microarray experiments were done to survey gene expression changes for the wild type, as compared to a mutant with the Lrp gene knocked-out (Tani, Khodursky, Blumenthal, Brown, and Matthews (2002)). It is intuitively reasonable that, in such experiments, the genes with expression changes are more likely to be regulated by the Lrp than those with no altered expression. Therefore, it is natural to conduct a combined analysis of DNA binding data and expression data.

A simple way is to analyze DNA binding data and expression data *separately* with each resulting in a list of genes, then take an intersection of the two lists and identify the common set of genes as the targets (Ren et al. (2000)). This approach can reduce the false positive number, but is likely to yield a high false negative number: many target genes may be missing from the common set; it is even possible that there is no intersection between the two lists of the genes. Existing integrated analysis strategies include regressing one source of data on the other (e.g., Conlon, Liu, Lieb and Liu (2003), Zhao, Wu and Sun (2003), Gao, Foat and Bussemaker (2004) and Sun, Carroll and Zhao (2006)), using one source to validate the other (e.g., von Mering et al. (2002)), sequential methods of using one source to generate hypotheses or priors for the following analysis on the second source of data (Liu, Brutlag and Liu (2002) and Xie, Pan, Jeong and Khodursky (2007)), and combining (e.g., taking an intersection of) the results of separate analyses on individual sources of data (Ren et al. (2000) and Xiao and Pan (2005)). A potentially more powerful, though more challenging, approach is to jointly model the two sources of data to improve the statistical power for new discoveries, as advocated and demonstrated by Holmes and Bruno (2000) for integrating DNA motif-finding and expression profile clustering.

Here we propose a novel joint model for protein-DNA binding data and gene expression data such that information in gene expression data is combined to nonparametrically infer the binding targets of a TF; to our knowledge, this is the first endeavor to do so in the literature. The basic idea is to exploit the correlation between TF binding and altered expression of a gene. We point out a connection of our proposal with the likelihood ratio test, thus establishing the optimality of our proposal. We extend a nonparametric empirical Bayes (EB) approach, proposed by Efron, Tibshirani, Storey and Tusher (2001) for gene expression data alone, to the joint model; in particular, we propose a novel conditional likelihood method to estimate two key mixing parameters for the expression data. That are not identifiable based on the expression data alone. We compare our proposal with other methods based on separate analyses of the two types of data: an intersection method, and an analog of Fisher's method to combine the results from separate analyses. Because the two sources of data support two different hypotheses, we clarify why such combined analyses are meaningful. Using both simulated and real data, we demonstrate that the joint modeling improves over the other methods.

2. Joint Modeling of Binding Data and Expression Data

2.1. Data and analysis goal

We assume throughout that we have DNA binding data of a TF (e.g., Lrp) as the primary data, with secondary data drawn from a gene expression experiment comparing a wild type against a mutant with partial or full loss of function of the TF (e.g., the deletion of the TF gene), both in the format of cDNA array data. Specifically, suppose that M_{1ij} and M_{2ij} are the log ratios of the intensities of the two channels for gene i on array j for the binding data and expression data, respectively, $i = 1, \ldots, G$, $j = 1, \ldots, n_1$, in the binding experiment, and $j = 1, \ldots, n_2$ in the expression experiment. Suppose that necessary data normalization has been accomplished. The central goal is to identify the genes bound by the TF with $E(M_{1ij}) \neq 0$. With the expression data alone, we can only detect differentially expressed (DE) genes with $E(M_{2ij}) \neq 0$. Note that the expression data cannot give unambiguous evidence for DNA-protein binding because a DE gene may or may not be bound by the TF.

For our purposes, for each gene i we construct two test statistics, X_i and Y_i based on the binding data and the expression data respectively. Although any test statistic can in principle be used, in this paper we consider the use of the SAM statistic, a regularized t-statistic (Tusher, Tibshirani and Chu (2001)), due to its simplicity and good performance (Xie, Jeong, Pan, Khodursky and Carlin (2004)). Specifically, suppose that the sample mean and the sample variance for gene i are $\bar{M}_{1i} = \sum_{j=1}^{n_1} M_{1ij}/n_1$ and $S_{1i}^2 = \sum_{j=1}^{n_1}(M_{1ij} - \bar{M}_{1i})^2/(n_1 - 1)$. Then $X_i = \bar{M}_{1i}/(S_{1i} + s_{10})$, where $s_{10} = \text{median}(S_{11}, \ldots, S_{1n_1})$ is used to stabilize the denominator. There are Bayesian justifications for the use of s_{10} (Baldi and Long (2001), Wright and Simon (2003) and Cui, Hwang, Qiu, Blades and Churchill (2004)). Similarly we define Y_i for $i = 1, \ldots, G$.

2.2. A joint model

Now we propose a joint model for (X_i, Y_i), the test statistics calculated from the two types of data, respectively. Define B_i as the indicator of whether gene i is indeed bound by the TF. Our goal is to identify all the genes (i.e., i's) with $B_i = 1$. We assume that $B_i \sim Bern(\pi)$, independently.

Our joint model consists of two mixture models for binding data and expression data, respectively. First, we specify a mixture model for the binding data:

$$f(X_i) = (1 - \pi)f_0(X_i) + \pi f_1(X_i), \tag{1}$$

where f_0 is the distribution of X_i for the genes with no binding, f_1 is that for the bound genes, and π is a prior probability of any gene's being bound by the TF. Second, we specify two mixture models for the conditional distribution of Y_i:

$$\begin{aligned} f(Y_i|B_i = 1) &= p_1 g_1(Y_i) + (1 - p_1)g_0(Y_i), \\ f(Y_i|B_i = 0) &= p_0 g_1(Y_i) + (1 - p_0)g_0(Y_i). \end{aligned} \tag{2}$$

This implies a two-component mixture model for the marginal distribution of Y_i:

$$g(Y_i) = (1-\pi_g)g_0(Y_i) + \pi_g g_1(Y_i), \tag{3}$$

where g_0 is the distribution of Y_i for genes with no expression changes, while g_1 is for genes with altered expression, and $\pi_g = \pi p_1 + (1-\pi)p_0$.

The main motivation of the model is the following. Intuitively, if a gene is regulated by the TF, there is a higher chance (probability p_1) that the gene is DE with a distribution g_1; otherwise, there is only a smaller chance (probability p_0) that the gene's expression level will be changed. The difference between p_1 and p_0 measures the amount of binding information contained in the expression data; in particular, if $p_1 = p_0$, then the binding and expression are completely independent. Note that, because binding is neither a sufficient nor a necessary condition for expression change, the above mixture model takes account of the possibility of a non-bound gene's having expression change.

We also assume that, conditional on the binding status B_i, a gene's binding statistic X_i and expression statistic Y_i are independent. Hence, the joint distribution of X_i and Y_i is

$$f(X_i, Y_i) = f(X_i|B_i = 1)f(Y_i|B_i = 1)\pi + f(X_i|B_i = 0)f(Y_i|B_i = 0)(1-\pi),$$

where $f(X_i|B_i = 1) = f_1(X_i)$ and $f(X_i|B_i = 0) = f_0(X_i)$. Using Bayes Theorem, we have the posterior probability of gene i being a target as

$$Pr(B_i = 1|X_i, Y_i) = \frac{\pi f(X_i|B_i = 1)f(Y_i|B_i = 1)}{f(X_i, Y_i)}, \tag{4}$$

which is used to draw inference on whether gene i is a binding site of the TF.

Remark 1. Optimality. From (3), we have

$$Pr(B_i\!=\!1|X_i, Y_i)\!=\!\frac{\pi}{\pi\!+\!(1\!-\!\pi)\frac{f(X_i|B_i=0)}{f(X_i|B_i=1)}\frac{f(Y_i|B_i=0)}{f(Y_i|B_i=1)}}\!=\!\frac{\pi}{\pi\!+\!\frac{1-\pi}{\mathrm{LRT}(B_i|X_i)\mathrm{LRT}(B_i|Y_i)}},$$

where $\mathrm{LRT}(B_i|X_i) = f(X_i|B_i = 1)/f(X_i|B_i = 0)$ and $\mathrm{LRT}(B_i|Y_i) = f(Y_i|B_i = 1)/f(Y_i|B_i = 0)$ are the two likelihood ratio test (LRT) statistics for testing H_{0i}: $B_i = 0$ vs H_{1i}: $B_i = 1$ based on the binding data X_i and expression data Y_i, respectively. On the other hand, the corresponding LRT statistic based on both the binding and expression data is

$$\mathrm{LRT}(B_i|X_i, Y_i) = \frac{f(X_i, Y_i|B_i = 1)}{f(X_i, Y_i|B_i = 0)} = \mathrm{LRT}(B_i|X_i)\mathrm{LRT}(B_i|Y_i),$$

with the last equality holding because of the conditional independence between X_i and Y_i. Thus, if $\pi < 1$, $Pr(B_i = 1|X_i, Y_i)$ is an increasing function of the joint likelihood ratio test statistic $\text{LRT}(B_i|X_i, Y_i)$, and thus is optimal.

Remark 2. Robustness. The joint modeling can automatically account for varying amounts of information contained in the two sources of data. For example, if there is almost no binding information contained in the expression data Y_i (or more generally in a subspace of Y), then we have $f(Y_i|B_i = 1) \approx f(Y_i|B_i = 0)$ and

$$Pr(B_i = 1|X_i, Y_i) \approx \frac{\pi f(X_i|B_i = 1)}{\pi f(X_i|B_i = 1) + (1-\pi) f(X_i|B_i = 0)} = Pr(B_i = 1|X_i),$$

which is equivalent to using the binding data X alone. Similarly, one uses only expression data if there is almost no information in the binding data (due to, e.g., a too small sample size). More generally, $Pr(B_1 = 1|X_i, Y_i)$ dictates a data-adaptive weighting on each source of data based on their relative information contents. Hence, our approach accounts for possibly different heterogeneity and specificity of multiple sources of data.

2.3. Statistical inference using empirical Bayes

Our joint model is general and flexible, allowing the unknown parameters in the model to be estimated by extending some existing methods for gene expression data to the current context. Here we consider a nonparametric EB method (Efron et al. (2001))

In each mixture model for one source of data, we use the observed data X and Y, respectively, to estimate f and g nonparametrically using finite Normal mixture models (Pan, Lin and Le (2003)), where the number of components is determined by a model selection criterion, such as BIC. As in Efron et al. (2001) and Pan et al. (2003), we permute the original binding data M_{1ij} and expression data M_{2ij} to estimate f_0 and g_0. We randomly keep or flip the sign of each M_{kij}, that is, $M^*_{kij} = a * M_{kij}$ with $a = 1$ or -1 with equal probability. Calculating the SAM statistics on the permuted data M^*_{kij}'s, we obtain permuted test statistics $X^* = (X^*_1, \ldots, X^*_{n_1})$ and $Y^* = (Y^*_1, \ldots, Y^*_{n_2})$. Again we fit Normal mixture models to X^* and Y^* to estimate f_0 and g_0, respectively.

As suggested in Efron et al. (2001), although π (or π_g) is not identifiable nonparametrically based on only X (or Y), a sensible estimate (more exactly, its lower bound) is

$$\hat{\pi} = 1 - \frac{\int_A \hat{f}(z)dz}{\int_A \hat{f}_0(z)dz} \qquad (5)$$

with A a small interval around 0; similarly for $\hat{\pi}_g$. Then estimates of f_1 and g_1 can be obtained: $\hat{f}_1 = \hat{f}/\hat{\pi} - \hat{f}_0(1-\hat{\pi})/\hat{\pi}$ and $\hat{g}_1 = \hat{g}/\hat{\pi}_g - \hat{g}_0(1-\hat{\pi}_g)/\hat{\pi}_g$.

The marginal distribution of Y_i, conditional on X_i, is

$$f(Y_i|X_i) = \sum_{k=0}^{1} f(Y_i|X_i, B_i = k) f(B_i = k|X_i) = \sum_{k=0}^{1} f(Y_i|B_i = k) Pr(B_i = k|X_i),$$

where the last equality follows from the conditional independence between X_i and Y_i. Hence, plugging-in the estimates and using (2) and (7) (see next), conditioning on X_i's, we obtain the maximum conditional likelihood estimates of p_1 and p_0,

$$(\hat{p}_1, \hat{p}_0) = \text{argmax}_{(p_1,p_0)} \prod_{i=1}^{G} \left\{ [p_1\hat{g}_1(Y_i) + (1-p_1)\hat{g}_0(Y_i)] \frac{\hat{\pi}\hat{f}_1(X_i)}{\hat{f}(X_i)} \right. \\ \left. + [p_0\hat{g}_1(Y_i) + (1-p_0)\hat{g}_0(Y_i)] \left(1 - \frac{\hat{\pi}\hat{f}_1(X_i)}{\hat{f}(X_i)}\right) \right\}.$$

Note that, without the binding data, the parameters p_1 and p_0 are not identifiable.

After obtaining the estimates of the parameters, we can plug them in to estimate the posterior probability $\widehat{Pr}(B_i = 1|X_i, Y_i)$ for each gene i, and then declare the genes with high $\widehat{Pr}(B_i = 1|X_i, Y_i)$ (i.e., larger than a cut-off value c) as the significant target genes. Similarly we can calculate $\widehat{Pr}(B_i = 1|X_i)$ and $\widehat{Pr}(DE_i|Y_i)$, thus giving two lists of the significant target genes. The cut-off value will be determined using the false discovery rate (FDR) (Benjamini and Hochberg (1995)) to be discussed next.

2.4. FDR and its estimation

It is important to estimate FDR for any given cut-off value c. Several methods have appeared to estimate FDR based on permutations (e.g., Efron et al. (2001), Xu, Olson and Zhao (2002), Pan (2003) and Storey and Tibshirani (2003)). Here we consider an approach based on a direct use of the posterior probability $Pr(B_i = 1|\cdot)$ that has been shown to work better than permutation-based methods (Newton, Noueiry, Sarkar and Ahlquist (2004)). Specifically, for any given cut-off value c, the corresponding FDR is

$$FDR(c) = \frac{\sum_{i=1}^{G} \beta_i I(\beta_i \le c)}{\sum_{i=1}^{G} I(\beta_i \le c)}, \qquad (6)$$

where $I()$ is an indicator function and $\beta_i = Pr(B_i = 0|\cdot) = 1 - Pr(B_i = 1|\cdot)$. Plugging in the parameter estimates, we obtain an estimated FDR.

3. Other Approaches

3.1. Separate analyses and combining their results

Based on the binding data X_i or expression data Y_i alone, respectively, we have the posterior probabilities

$$Pr(B_i = 1|X_i) = \frac{\pi f_1(X_i)}{f(X_i)}, \qquad Pr(DE_i|Y_i) = \frac{\pi_g g_1(Y_i)}{g(Y_i)}, \tag{7}$$

and use them to infer whether gene i is a binding target. Note that, because a DE gene may or may not be bound by the TF, DE genes are used in practice only as putative binding targets (e.g., Tani et al. (2002)), and, in general, incorrectly so.

A simple method to combine the above two separate analyses is to take the intersection of their identified gene lists; equivalently, we define

$$\widetilde{Pr}(B_i = 1|\cdot) = \min\{Pr(B_i = 1|X_i), Pr(DE_i|Y_i)\}$$

and compare $\widetilde{Pr}(B_i = 1|\cdot)$ to a cut-off value to declare significant genes. The intersection method is simple and intuitively reasonable, and is used in practice (Ren et al. (2000)).

Another simple method, analogous to Fisher's method of combining two p-values, defines

$$\widetilde{\widetilde{Pr}}(B_i = 1|\cdot) = \sqrt{Pr(B_i = 1|X_i)Pr(DE_i|Y_i)}$$

and then uses this to select significant genes. We call it Fisher's method.

3.2. A justification

Combining separate analyses is intuitively reasonable, but ad hoc; in particular, because $Pr(DE_i|Y_i)$ provides statistical evidence for DE, not for binding as supported by $Pr(B_i = 1|X_i)$, it is not immediately clear why this type of method would work. Below, based on our mixture models, we provide a justification which, along with Remark 1 in Section 2.2, also explains why these methods are suboptimal as compared to the joint modeling.

It is easy to see that

$$Pr(B_i = 1|X_i) = \frac{\pi}{\pi + \frac{(1-\pi)}{\mathrm{LRT}(B_i|X_i)}}, \quad Pr(DE_i|Y_i) = \frac{\pi_g}{\pi_g + \frac{(1-\pi_g)}{\mathrm{LRT}(DE_i|Y_i)}},$$

where $\mathrm{LRT}(DE_i|Y_i) = g_1(Y_i)/g_0(Y_i)$ is the likelihood ratio statistic to test for DE. On the other hand,

$$\mathrm{LRT}(B_i|Y_i) = \frac{p_1 g_1(Y_i) + (1-p_1)g_0(Y_i)}{p_0 g_1(Y_i) + (1-p_0)g_0(Y_i)} = \frac{p_1 \mathrm{LRT}(DE_i|Y_i) + 1 - p_1}{p_0 \mathrm{LRT}(DE_i|Y_i) + 1 - p_0},$$

which is an increasing function of $\mathrm{LRT}(DE_i|Y_i)$ if $p_1 > p_0$, as expected. Thus, a method combining the two separate posterior probabilities $Pr(B_i = 1|X_i)$ and $Pr(DE_i|Y_i)$ is equivalent to combining the two likelihood ratio statistics based on each data source alone.

3.3. Other methods

In the Supplement, we introduce two special cases of the joint modeling with possibly over-simplified assumptions, and a new sequential Bayesian method that uses the gene expression data to generate priors for the subsequent analysis of binding data. It turns that the sequential method is related to a special case of the joint modeling, and none worked as well as the joint modeling.

4. Simulation

To demonstrate the feasibility and potential gain of our proposal, we did a simulation study.

4.1 Simulation set-ups

To be as practical as possible, we generated simulated data by mimicking the Lrp binding data and expression data. First, based on the actual Lrp binding data, we calculated the SAM statistic (Tusher et al. (2001)) for each gene. Then we picked up the top G_1 genes with the largest SAM statistics and treated them as the true binding targets of Lrp. For each of the G_1 targets, we simulated its binding log-ratios from a Normal distribution with mean and variance the sample mean and sample variance from the data; for the other $(4,281 - G_1)$ non-target genes, their log-ratios were generated independently from Normal distributions with mean 0 and variances equal to their sample variances in the binding data.

To generate a expression data set, we first randomly selected $p_1 G_1$ genes from the G_1 target genes as DE genes; second, among the other $(4,281 - G_1)$ non-target genes, we randomly selected a proportion p_0 of them as DE genes, and the remaining ones as equally-expressed (EE) genes. Again the expression levels of each gene were simulated from a Normal distribution with variance equal to its sample variance in the expression data, and mean 0 if it was selected as an EE gene, or mean equal to its sample mean in the data if it was selected as a DE gene.

To mimic the actual data, we had total 4,281 genes, 5 replicates for the binding data and 6 replicates for the expression data. The simulation was conducted in R (Ihaka and Gentleman (1996)). In particular, we used package `mclust` to fit a finite Normal mixture model with the number of components selected by BIC (Fraley and Raftery (2003)).

Three sets of parameter values were used: (i) $G_1 = 400$ (and thus $\pi = 400/4281 = 0.093$), $p_1 = 0.9$ and $p_0 = 0.2$ (and thus $\pi_g = \pi p_1 + (1-\pi)p_0 = 0.265$); (ii) $G_1 = 400$ and $p_1 = p_0 = 0.2$; (iii) similar to Case (i) except that the replicates of both binding and expression data were correlated: there was a within-gene (i.e., between-array) correlation of 0.1 for each gene; for a dataset with eight arrays, Efron (2004, Table 3) demonstrated that there were positive pairwise correlations among the first four arrays and the next four, with a median of 0.085, which motivated our choosing the within-gene correlation at 0.1. Specifically, for Case (iii), if a binding log-ratio $M'_{ij} \sim N(\mu_{x,i}, \sigma^2_{x,i})$, a random effect $b_i \sim N(0, \sigma^2_{x,i})$, and if M'_{ij} and b_i are independent, then $M_{ij} = M'_{ij} + b_i$ was gene i's binding log-ratio on array j; expression data were generated similarly. In Case i), with $p_1 > p_0$, the expression data contained some information about binding. Case (ii) represented a null case where, due to $p_1 = p_0$, there was no information about binding contained in the expression data. We considered the robustness of the methods in Case (iii), where the commonly used assumption of independent arrays was incorrect, which could happen in practice (Efron (2004)). Note that, although our methods are all nonparametric without strong distributional assumptions, a modeling assumption was violated even in Cases (i) and (ii): the null distribution of the SAM statistics was not a finite mixture of normals as used in our estimation procedure. The goal of simulation was to show that a joint analysis could improve over using binding data alone if expression data indeed contained binding information and, at the same time, the joint analysis did not deteriorate otherwise.

Two more simulation set-ups are considered in the Supplement.

4.2. Results

The averages of the estimates based on 100 simulations for each set-up are summarized in Figure 1. Both a receiver operating characteristic (ROC) plot and a realized FDR plot was used to compare the performance of the four methods: using the binding data alone, combining the separate analyses of binding and expression data using the intersection method or Fisher's method, and the joint analysis. For any simulated dataset with cutoff c, the claimed positive number is $\widehat{TP}(c) = \#\{i : Pr(B_i = 1|\cdot) \geq c\}$, which is a sum of true positive number TP and false positive number FP:

$$TP(c) = \#\{i : Pr(B_i = 1|\cdot) \geq c, B_i = 1\},$$
$$FP(c) = \#\{i : Pr(B_i = 1|\cdot) \geq c, B_i = 0\}.$$

The realized sensitivity, specificity and FDR are

$$sens(c) = \frac{TP(c)}{G_1}, \quad spec(c) = 1 - \frac{FP}{(G - G_1)}, \quad FDR(c) = \frac{TP(c)}{\widehat{TP}(c)},$$

while the estimated FDR was obtained from (6); their averages across 100 simulations are plotted in Figure 1.

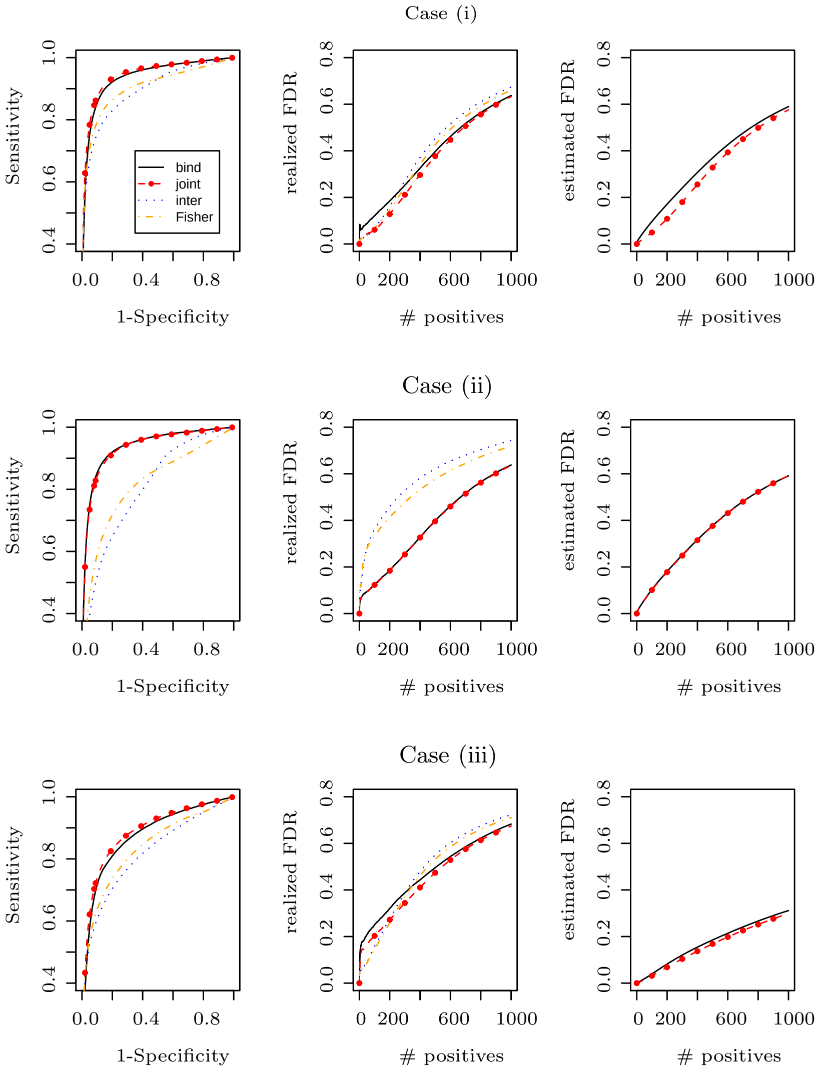

Figure 1. ROC curves, realized/estimated FDR vs number of estimated significant genes for various methods for simulated data. In Case (ii), the curves for “bind” (i.e., using binding data alone) and “joint” (i.e., the joint model) completely overlap.

In Case (i) with binding information contained in expression data, joint analysis was the obvious winner: it was the most powerful with the highest sensitivity for a given specificity. In particular, compared to using binding data alone and at a high specificity level, the joint analysis shows a large improvement in sensitivity over using binding data alone; consistently, at a low FDR, the joint analysis gave many more positives (Table 1). Notably, this efficiency gain of joint analysis was achieved even though the information content in the expression data was limited: among all the genes with expression changes, $(1-\pi)p_0/\pi_g = 68\%$ were not binding targets. In general, the information content contained in expression data, or the degree of correlation between the two types of data, can be measured by $(1-\pi)p_0/\pi_g$ or, for a fixed π, by the ratio p_1/p_0; it was confirmed that there was even a larger efficiency gain from the joint analysis as p_1/p_0 increased (results not shown). The simulation set-up of Case (i) was chosen to give a realistic scenario. On the other hand, the two methods of combining separate analyses performed well only if the number of the declared significant genes was small, and overall, surprisingly, they might not improve over analyzing binding data alone.

Table 1. Comparison of statistical powers of the joint analysis (joint) and using binding data alone (bind) with simulated data for Case (i): sensitivity vs specificity, and the claimed positive number ($\widehat{TP}$) vs realized FDR, all averaged over 100 simulations; Impr=100(joint-bind)/bind, the percentage improvement of the joint analysis over using binding data alone.

	spec	0.99	0.98	0.97	0.96	0.95	0.90	0.85	0.80
sens	binding	0.417	0.565	0.652	0.708	0.751	0.853	0.899	0.922
	joint	0.504	0.628	0.701	0.748	0.784	0.874	0.912	0.933
	Impr(%)	20.9	11.2	7.5	5.6	4.4	2.5	1.4	1.2
	FDR	0.05	0.06	0.07	0.08	0.09	0.10	0.15	0.20
$\widehat{TP}$	binding	1	8	23	37	54	65	146	222
	joint	74	98	117	134	147	162	239	288
	Impr(%)	7300	1125	409	262	172	149	64	30

When expression data contained no binding information, as in Case (ii), the joint analysis reduced to using the binding data alone: averaged ROC curves and realized FDR curves, respectively, almost completely overlapped with each other. It was reassuring that the joint analysis did not lose efficiency in a null case. In contrast, the two methods of combining separate analyses deteriorated dramatically. When replicated arrays were not independent, as in Case (iii), the same conclusions held as in Case (i), though all the methods performed worse than they did in Case (i). In particular, the joint analysis still performed best.

As Newton et al. (2004) pointed out, the FDR estimation depends on the adequacy of the fitted model. For example, using expression data alone, the posterior probabilities are for detecting DE genes, not for binding targets, thus the estimated FDRs are for detecting DE genes, not for binding targets; for this reason, the FDR estimates from the intersection and Fisher's methods were biased and are not plotted in Figure 1. For the other two methods, the FDR estimates were almost unbiased in Cases (i) and (ii); however, when the arrays were correlated as in Case (iii), the FDR estimates were under-biased: the two means of the FDR estimates for the two methods were smaller than their counterparts from the realized FDRs, see the two panels of Figure 1. A use of FDR estimates is to aid in choosing between two competing methods (Xie et al. (2004); for this purpose, albeit biased, the FDR estimates still gave the correct choice between the joint analysis and analyzing the binding data.

Table 2. Means and standard errors (SEs) of the mixing probability estimates from 100 simulations.

	Case (i)				Case (ii)				Case (iii)			
	π	π_e	p_1	p_0	π	π_e	p_1	p_0	π	π_e	p_1	p_0
true	0.093	0.265	0.900	0.200	0.093	0.200	0.200	0.200	0.093	0.265	0.900	0.200
mean	0.096	0.260	0.775	0.186	0.093	0.202	0.239	0.193	0.277	0.422	0.609	0.338
SE	0.003	0.003	0.005	0.002	0.003	0.003	0.004	0.002	0.003	0.002	0.004	0.002

The mean estimates of the four mixing parameters were given in Table 2; we used $A = [-0.05, 0.05]$ in (5) for both π and π_e. For Case (i) or (ii), π and π_e were estimated surprisingly well, though theoretically they were not identifiable; the estimate for p_0 was almost unbiased, while that for p_1 was biased, for which we have no explanation. For Case (iii), because the independence assumption was violated, as expected, the permutation method being used under-estimated f_0 and g_0 (Efron (2004)), leading to the over-biased estimates of π and π_e; these biased estimates in turn introduced biases for the estimates of p_1 and p_0.

5. Application to Lrp Data

We analyzed the Lrp data to identify the binding targets of Lrp. The binding data were generated in house from a ChIP microarray experiment. Briefly, DNA samples from wild type *Escherichia coli* were labelled with red (Cy5) fluorophore after crosslinks, immunoprecipitation and amplification, whereas genomic DNA samples were prepared and labelled with green (Cy3) fluorophore to serve as controls. It was intended to identify the binding locations of Lrp by comparative hybridization of the two samples to a DNA microarray. The Cy5 and Cy3 intensities at each spot on the array measured the relative abundances of the DNA subsequences bound by Lrp in the two samples, respectively. Hence the log-ratio of Cy5 to Cy3 intensities at each spot provided a measure of the extent of binding of Lrp to the corresponding genomic locus.

Tani et al. (2002) published a study using cDNA microarrays to survey gene expression changes between the cell of the wild type and that of a mutant with the gene encoding Lrp knocked out. Due to the obvious connection of this study with our Lrp binding experiment, we aim to borrow information from this gene expression dataset to help identify the binding targets of Lrp.

After combining the two datasets, we had in total 4,281 genes (ORFs). There were five replicates/arrays for the binding data, and six replicates for the gene expression data. Because of the use of genomic DNA as control samples in the ChIP experiment, we used a global normalization method; that is, we centered the log-ratios on each array at median 0 and scaled them by the inter-quartile range on the array. For the expression data, we took the standard local normalization using the loess smoother (Yang, Dudoit, Luu and Speed (2002)).

Figure 2 (top panel) gives the scatter plot of the test statistics of the binding data versus that of the expression data. There seems to be little marginal correlation between the two sets of the statistics; this could be due to the high noise level in either data source, or to the fact that there were many downstream genes indirectly regulated by Lrp. This attested to the challenge that this particular problem brought with only a limited amount of binding information contained in the expression data. Based on the performance of the methods in simulations, we only considered analyzing the binding data alone and the joint analysis.

Using the binding data alone, the posterior probability is a function of the binding test statistic (bottom panel, Figure 2). However, the posterior probability in the joint analysis is a function of both the binding test statistic and the expression statistic. For example, when the binding statistic is about 0.5, using the binding data alone gives a posterior probability around 0.6; however, depending on the expression statistic, the posterior probability from the joint modeling ranges from 0.4 to 0.7; as to be discussed next, this difference could lead to the joint analysis's identifying a known target (Lrp gene) that using the binding data alone missed.

Some parameter estimates were $\hat{\pi} = 0.185$, $\hat{p}_1 = 0.83$ and $\hat{p}_0 = 0.67$. Again the relatively small ratio between $\hat{p}_1$ and $\hat{p}_0$ suggested little information contained in the expression data. Nevertheless, based on the estimated FDR curves, it appears that the joint modeling reduced the FDR when compared with analyzing the binding data alone; see the Supplement.

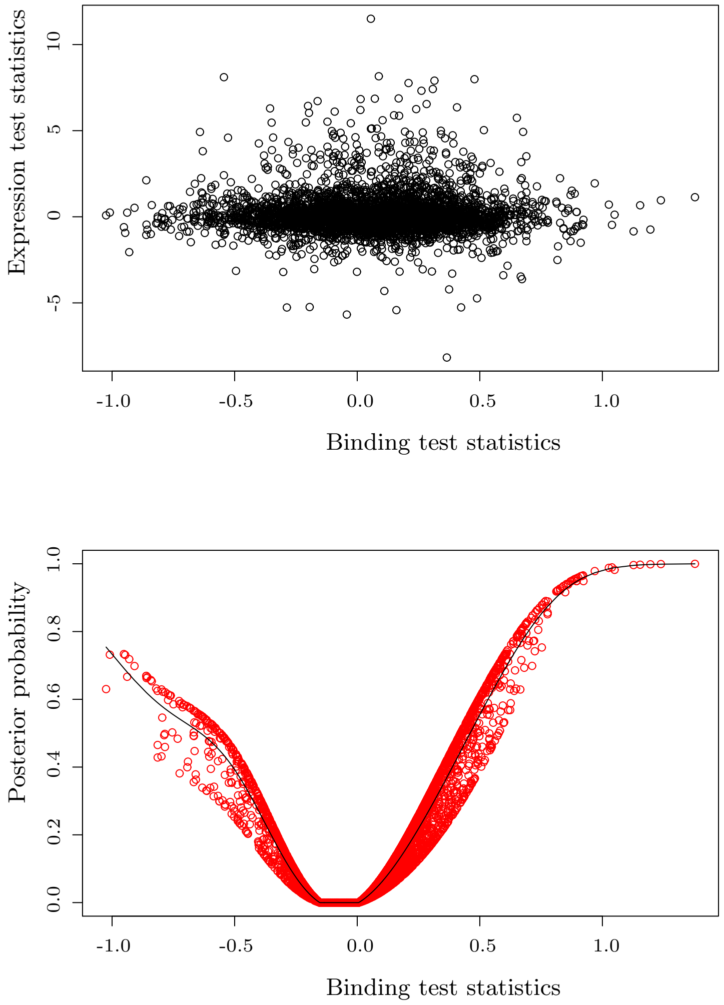

Figure 2. The scatterplot of the test statistics of the binding data vs. that of the expression data (top panel), and the estimated posterior probabilities using the binding data alone (solid line) and using the joint analysis (circles) (bottom panel) for the Lrp data.

We also conducted a biological evaluation based on a comprehensive literature search. Table 3 lists the genes/operons known to be bound by Lrp as discussed in the literature, their estimated posterior probabilities of being bound by Lrp, and the ranks of their posterior probabilities among all the genes, given by the analysis using the binding data alone and the joint analysis. For an operon containing more than one gene, for each method, we gave the maximum of the posterior probabilities (and thus the minimum of the ranks) among the genes in the operon. It can be seen that, for most of the genes/operons, the joint analysis gave higher posterior probabilities and higher ranks than did the analysis using the binding data alone. There were three exceptions: for operon *dadAX*, the joint analysis gave a slightly lower probability and a lower rank, largely due to the lack of evidence to support that either of the two genes, *dadA* and *dadX*, had altered expression: their estimated posterior probabilities of expression changes were ranked 2,789 and 3,487 respectively, while those of most other genes in Table 1 were ranked much higher; for genes *clpB* and *aidB*, both methods gave an estimated posterior probability of about 0. Hence, evidently, the joint model had an efficiency-gain over using the binding data alone: for example, if we used the usual cut-off at $c = 0.5$ (with estimated FDR about 0.35, see Figure 2 in the Supplement), the joint analysis was able to correctly identify gene *lrp* as a target (Wang et al. (1994) and Oshima et al. (1995)), whereas using the binding data alone would miss it.

Table 3. Genes known to be bound by Lrp as discussed in the literature, their posterior probabilities and ranks as being bound by Lrp using the binding data alone, and the joint analysis of the binding data and expression data.

	binding		joint	
Gene	Prob	Rank	Prob	Rank
serA	0.772	79	0.808	68
osmY	0.749	98	0.787	83
fimBE	0.674	140	0.719	117
ilvGMEDA	0.524	342	0.575	268
lrp	0.449	543	0.502	419
atpAD	0.430	597	0.483	473
gltBDF	0.399	697	0.451	558
dadAX	0.386	738	0.380	751
ilvIH	0.265	1194	0.309	1001
lysU	0.148	1902	0.177	1679
csiD	0.132	2012	0.158	1798
gcvTHP	0.130	2023	0.156	1812
osmC	0.030	3018	0.037	2894
ompC	0.019	3155	0.023	3061
tdh	0.006	3364	0.008	3318
clpB	0.000	-	0.000	-
aidB	0.000	-	0.000	-

6. Discussion

With the rapid accumulation of various high-throughput genomic and proteomic data, there is an increasing interest to develop new statistical methods that can take advantage of the existence of multiple types and sources of data. However, we are not aware of any other existing work aiming to jointly model DNA-protein binding data and expression data in spite of the significant use of the two types of data in wide-ranging applications, and our proposal can be also applied to other and more than two sources of genomic data. Among the existing approaches, the closest to ours was proposed by Wang et al. (2005), in which a similar joint model was used for DNA sequence data and either gene expression data or binding data; however, in addition to different data sources, a key difference is that they used a parametric model. Bar-Joseph et al. (2003) also considered combining information from binding and expression data to infer a common set of target genes for a group of TFs: first, they used the binding data with a stringent cutoff to obtain an initial list L_1 of the target genes for a group T of TFs; second, they used the expression profiles across multiple experimental conditions to obtain another set L_2 of the genes that were strongly co-expressed with the genes in L_1; third, they used the binding data with a less stringent cutoff to add a possible subset of L_2 into L_1, which was taken as the output. In particular, Fisher's method was used in the last step to combine the p-values for a gene in L_2 to be bound by the TFs in T; each p-value was obtained from the binding data alone. Hence, their method was more of a sequential strategy for using expression data to generate priors for analyzing binding data (thus relaxing the cutoff to identify targets based on the binding data), as in Xie, Pan, Jeong and Khodursky (2007).

Here we have proposed a novel joint model to nonparametrically analyze the two types of data simultaneously to identify binding targets of a TF. The basic idea is to exploit the correlation between the TF binding and expression change of a gene, thus enabling borrowing information from expression data to detect binding targets. We have demonstrated the feasibility as well as possible efficiency gain of the joint modeling over several existing methods. In our motivating example, as a broad transcription regulator, Lrp binds to relatively a large number of genes, some of which further regulate many other genes' expression; in other words, there are probably many downstream genes that are indirectly regulated by Lrp. Therefore, there is only a limited information content on binding contained in the expression data, leading to only moderate improvement of the results in the joint analysis. Nevertheless, because of the prior existence of the expression data, it is still desirable to have a joint analysis: at no extra experimental cost, it resulted in more biologically confirmed binding targets as demonstrated in our example. Furthermore, it is conceivable that for other less general TFs with less downstream genes indirectly regulated, the joint analysis will result in a larger efficiency gain. Finally, a nice property of the joint model is its robustness: if there is indeed no binding information contained in the expression or other secondary data, the joint model reduces to analyzing the binding data alone.

In summary, a main message of this study is that the joint analysis improves over analyzing a single data source and other ad hoc methods of combining two separate analyses, even when the secondary data may contain only a limited amount of information. Presumably, incorporation of other sources of data, such as DNA sequence data, into the above joint modeling framework will further improve efficiency gain. This is currently under investigation.

We comment on two issues related to biology. First, there is a recent technological innovation in using tiling arrays to map DNA-protein binding locations. A feature is that there are several probes (i.e., DNA subsequences) corresponding to each gene on an array. A common theme of existing approaches to analyzing tiling arrays is to smooth the signals of neighboring probes with a sliding window on a chromosome, and then to identify a signal peak for each gene, resulting in a summary expression or binding statistic for each gene (Buck, Nobel and Lieb (2005) and Ji and Wong (2005)). In this way, with a summary statistic for each gene, our joint model can be directly applied. Second, in the present study, the expression experiment surveyed the expression difference between the wild type and a strain with the gene encoding a TF knocked out. There might be a concern on the availability of such deletion experiments, but they are not necessary; for example, it is appropriate to use any data with a partial or full loss of the TF function, including deletions, conditional mutations and data obtained by RNA interference (RNAi).

In this work, we have used a model-based FDR estimation procedure proposed by Newton et al. (2004) and found that it worked better than permutation-based methods, in agreement with the conclusion found there. In particular, the flexibility of the joint model enables the use of model-based FDR estimation, in contrast to the problems associated with other *ad hoc* methods, such as the intersection method, due to their questionable modeling assumptions. Although it is conceptually possible to use other estimation methods, we have adopted a nonparametric EB approach of analyzing gene expression data (Efron et al. (2001)) to the current context for statistical inference. The nonparametric EB approach is particularly attractive with regard to its flexibility and simplicity. However, there is room for improvement. First, the prior probability in the nonparametric mixture model for the primary data is not identifiable (Efron et al. (2001)); we used a simple estimate of Efron et al. (2001), and other more sophisticated estimators may be also used (e.g., Storey and Tibshirani (2003), Pounds and Cheng (2004) and Dalmasso, Broet and Moreau (2005)). Second, there may be problems with permutation-based methods to estimate the null distribution (i.e., f_0 or g_0) (Pan (2003) and Efron (2004)); other empirical estimates may be applied (Efron (2004) and McLachlan, Bean and Jones (2006)). Alternatively, one may try parametric EB (Newton et al. (2001) and Kendziorski et al. (2003)), semi-parametric EB

(Newton et al. (2004)) or fully Bayesian approaches (Do, Muller and Tang (2005) and Lewin et al. (2006)). These are currently under investigation.

Acknowledgements

We are most grateful to an AE and two reviewers for many helpful and constructive comments. We thank Dr. Rowena Matthews from the University of Michigan in Ann Arbor for providing us with Lrp antibodies. This research was supported by NIH grants HL65462 (WP and YX) and GM066098 (KSJ, YX and AK), and a UM AHC Faculty development grant (WP and AK).

References

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* **21**, 1337-1342.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Buck, M. J., Nobel, A. B. and Lieb, J. D. (2005). ChIPOTle: a use friendly tool for the analysis of ChIP-chip data. *Genome Biology* **6**:R97.

Conlon, E. M., Liu, X. S., Lieb, J. D. and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. USA* **100**, 3339-3344.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J. and Churchill, C. R. (2004). Improved statistical tests for differential gene expression by shrinking variance components estimates. To appear in *Biostatistics*.

Dalmasso, C., Broet, P. and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *Bioinformatics* **21**, 660-668.

Do, K.-A., Muller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Appl. Statist.* **54**, 627-644.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96-104.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. G. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151-1160.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, discriminant analysis, and density estimation: MCLUST. *J. Classification* **20**, 263-286.

Gao, F., Foat, B. C. and Bussemaker, H. J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**:31.

Holmes, I. and Bruno, W. J. (2000). Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 202-210.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5**, 299-314.

Ji, H. and Wong, W. H. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* **21**, 3629-3636.

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statist. Medicine* **22**, 3899-3914.

Lee, M.-L. T., Bulyk, M. L., Whitmore, G. A. and Church, G. M. (2002). A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics* **58**, 981-988.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006). Bayesian modelling of differential gene expression. *Biometrics* **62**, 1-9.

Liu, J. S., Neuwald, A. F. and Lawrence, C. E. (1999). Markovian structures in biological sequence alignments. *J. Amer. Statist. Assoc.* **94**, 1-15.

Liu, X. S., Brutlag, D. L. and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835-839.

McLachlan, G.J., Bean, R.W., Jones, L.B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608-1615.

Newton, M. A, Kendziorski, C. M., Richmond, C. C., Blattner, F.R. and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Computational Biology* **8**, 37-52.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155-176.

Oshima, T., Ito, K., Kabayama, H. and Nakamura, Y. (1995). Regulation of lrp gene expression by H-NS and Lrp proteins in Escherichia coli: dominant negative mutations in lrp. *Mol Gen Genet.* **247**, 521-528.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **12**, 546-554.

Pan, W. (2003). On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* **19**, 1333-1340.

Pan, W., Lin, J. and Le, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics* **3**, 117-124.

Pounds, S. and Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737-1745.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-2309.

Smyth, G. K., Yang, Y. H. and Speed, T. (2003). Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols* (Edited by M. J. Brownstein and A. B. Khodursky), 111-136, Methods in Molecular Biology Volume 224, Humana Press, Totowa, NJ.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide experiments. *Proc. Natl Acad. Sci. USA* **100**, 9440-9445.

Sun, N., Carroll, R.J. and Zhao, H. (2006). Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proc. Natl. Acad. Sci. USA* **103**, 7988-7993.

Tani, T., Khodursky, A., Blumenthal, R., Brown, P., and Matthews, R. (2002). Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis. *Proc. Natl. Acad. Sci. USA* **99**, 13471-13476.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci. USA* **98**, 5116-5121.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.

Wang, Q., Wu, J., Friedberg, D., Plakto, J. and Calvo, J. M. (1994). Regulation of the Escherichia coli lrp gene. *J. Bacteriol.* **176**, 1831-1839.

Wang, W., Cherry, J. M., Nochomovitz, Y., Jolly, E., Botstein, D. and Li, H. (2005). Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc. Nat. Acad. Sci. USA **102**, 1998-2003.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448-2455.

Xiao, G. and Pan, W. (2005). Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data. *J. Bioinformatics and Computational Biology* **3**, 1371-1389.

Xie, Y., Jeong, K. S., Pan, W., Khodursky, A. and Carlin, B. P. (2004). A case study on choosing normalization methods and test statistics for microarray data. *Comparative and Functional Genomics* **5**, 432-444.

Xie, Y., Pan, W., Jeong, K. S. and Khodursky, A. (2007). Incorporating prior information via shrinkage: a combined analysis of genome-wide location data and gene expression data. *Statist. Medicine* **26**, 2258-2275.

Xu, X. L., Olson, J. M. and Zhao, L. P. (2002). A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Human Molecular Genetics* **11**, 1977-1985.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. (2002). Normalization for cDNA Microarray Data. *Nucleic Acids Research* **30**, e15.

Zhao, H., Wu, B. and Sun, N. (2003). DNA-protein binding and gene expression patterns. In *Science and Statistics: A Festschrift for Terry Speed* (Edited by D. R. Goldstein), 259-274. IMS Lecture Notes-Monograph Series 40.

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455, U.S.A.

E-mail: weip@biostat.umn.edu

Department of Biological Chemistry, UCLA School of Medicine and Molecular Biology Institute, University of California, Los Angeles, 611 Charles E. Young Dr. East, Boyer Hall Rm320, Los Angeles, CA 90095-0001, U.S.A.

E-mail: jeongks@ucla.edu

Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd.

E-mail: yang.xie@utsouthwestern.edu

Department Biochemistry, Molecular Biology and Biophysics, University of Minnesota, 6-155 Jackson Hall 321 Church St. Minneapolis, MN 55455, U.S.A.

E-mail: khodu001@umn.edu

(Received May 2006; accepted November 2006)