DECOMPOSING GENE EXPRESSION INTO REGULATORY AND DIFFERENTIAL PARTS WITH BAYESIAN DATA FUSION

Janne Nikkilä[1]*, Timo Erkkilä*[2] *and Harri Lähdesmäki*[2]

[1]Adaptive Informatics Research Centre & Helsinki Institute for Information Technology
Department of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland
janne.nikkila@tkk.fi
[2]Department of Signal Processing
Tampere University of Technology
P.O. Box 553, FI-33101 Tampere, Finland
{timo.p.erkkila,harri.lahdesmaki}@tut.fi

ABSTRACT

The two main interests in gene expression data, differential expression and transcriptional regulatory effects, are usually difficult to separate from each other. We propose a method for decomposing observed gene expression data into i) a part explainable directly by transcription factor (TF) mRNA level, and ii) a part attributable to other effects induced by experimental setting. The method fits a Bayesian hierarchical linear model to the expression data given prior information about transcriptional regulatory mechanisms. Our primary source of prior information are TF binding probabilities, derived from a probabilistic model for TF binding to gene regulatory sequences. The proposed method can be easily extended to include additional and other types of prior information (ChIP-chip, other gene expression data), and the same modeling framework can be used to make inference regarding a large variety of questions. Simulation results show that, relative to standard approaches, the proposed method can better detect regulatory relations and that it is also able to distinguish general differential expression from the effects of direct regulatory mechanisms.

1. INTRODUCTION

Detection of differential gene expression induced by the chosen experimental setting is the primary focus in most microarray gene expression studies. In addition to pure differential expression, key interests are the gene regulation effects taking place during the experiment. While gene regulation can happen in various stages, one of the most important of these stages is direct transcriptional regulation by transcription factor (TF) proteins. Approximate protein levels of TFs can in principle be monitored through the expression levels of the genes coding the factor proteins. However, estimating these regulatory effects from gene expression data is not a trivial task. The most important challenges include problems in estimation due to small number of samples, regulatory relations is confounded with co-expression of genes and TFs, and differential expression of a gene might be induced by other TFs than the known regulating TFs.

Discovering regulatory relations from high-throughput gene expression data has been in focus since the emergence of microarrays. The earliest and most common attempts for finding relations between genes were based on detecting genes with similar behaviour in the experiments [1]. Naturally, this results in a set of genes consisting both the regulators, the target genes, and the co-expressed genes with no means to distinguish between them. More focused approaches have been presented as well which rely on prior knowledge about the potential regulators. For example, Bayesian networks have also been applied to estimate regulatory relations between a set of known TFs and the rest of the genes as groups (see [2]). However, practically all these approaches by-pass the actual interest of a single microarray gene expression experiment: differential expression in the given situation.

Statistical models for differential expression range from early fold-change based approaches to classical ANOVA based models, and their Bayesian variants (see [3, 4]). But, analogously to regulation models, differential expression models largely disregard the estimation of the regulatory effects.

We propose here a model that integrates gene expression data with transcriptional regulatory knowledge, such as transcription factor binding site location information. Using the binding probabilities from a probabilistic model for transcription factor binding to gene promoter region as a prior, the proposed model can in principle distinguish which genes are transcriptionally regulated by known TFs in any given experiment, based on the observed expression data. In particular, the model is capable of distinguishing whether some known, differentially expressed TF is causing its target genes to be differentially expressed in the

given experiment. In addition, the model is able to discover differential expression of target genes due to other reasons than the known TF, in case the regulatory TF is not differentially expressed. The model is formulated in Bayesian framework enabling a natural way to handle the uncertainty in the data and small sample size problem.

The results show that the proposed model can detect regulatory relations taking place during the experiment more efficiently than mere co-expression based approaches, and at the same time detect differential expression induced by the experiment.

2. METHODS

Our approach is closely related to a Bayesian hierarchical model for gene expression allowing heterogenous errors (HEM) [4]. HEM separates the technical noise from the biological noise, and is shown the perform favorably in the analysis of gene expression data. We extend the standard HEM by incorporating an additional regression term that allows explicit modeling of direct transcriptional regulation. Further, the regression term allows to incorporate prior knowledge about transcriptional regulation into the hierarchical error model. The prior information can come from a variety of different sources, such as sequence-based TF binding predictions [5], ChIP-chip data [6], other gene expression data.

We propose to model the observed expression $y_{i,j,k}$ for ith gene, jth condition, and kth replicate with a linear model as

$$y_{i,j,k} = \mu + g_i + d_j + r_{i,j} + a_i \cdot z_i \cdot x_{TF,j} + \epsilon_{i,j,k}, \quad (1)$$

where μ is the general mean, g_i is the effect of ith gene, d_j is the effect of the jth condition, $r_{i,j}$ is their joint effect, $z_i \cdot a_i \cdot x_{TF,j}$ describes the regulatory effect of the given TF with TF's expression level $x_{TF,j} = \mu + g_{TF} + d_j + r_{TF,j}$, regulation strength a_i and a binary indicator z_i of whether the TF regulates the ith gene. The residual variance is described with $\epsilon_{i,j,k}$. For the expression measurements of the TF we use the same model but without the regression term, i.e. $y_{TF,j,k} = x_{TF,j} + \epsilon_{i,j,k}$.

We assume the following prior distributions with fixed parameters:

$$\begin{aligned}
\mu &\sim N(\mu_\mu, \sigma_\mu^2) && (2)\\
g_i, g_{TF} &\sim N(\mu_g, \sigma_g^2) && (3)\\
d_j &\sim N(\mu_d, \sigma_d^2) && (4)\\
r_{i,j}, r_{TF,j} &\sim N(\mu_r, \sigma_r^2) && (5)\\
a_i &\sim N(\mu_a, \sigma_a^2) && (6)\\
z_i &\sim Bernoulli(\theta_i) && (7)\\
\epsilon_{i,j,k} &\sim N(0, \tau_{i,j}^2) && (8)\\
\tau_{i,j}^{-2}, \tau_{TF,j}^{-2} &\sim Gamma(\alpha, \beta). && (9)
\end{aligned}$$

Note that the error variance is allowed to be heterogeneous, i.e. different for each gene and condition. Prior information about transcriptional regulation can easily be incorporated via θ_i parameters. The proposed model is able to analyze only one TF at a time. The graphical model is presented in Figure 1.

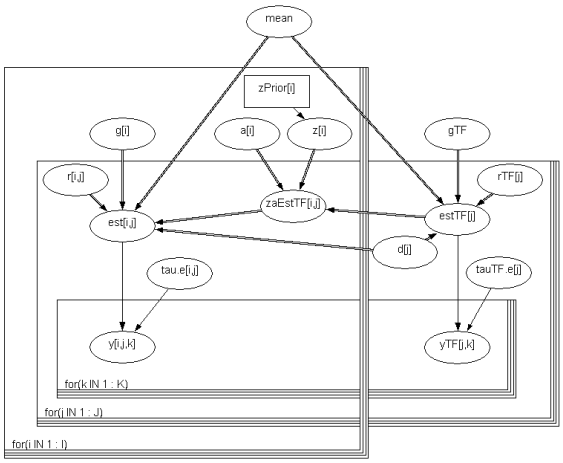


Figure 1. A graphical representation of the proposed model. The boxes indicate loops over samples i, conditions j, and replicates k.

The unknown model parameters are estimated with Gibbs sampling in WinBUGS [7]. Convergence to the posterior is assessed using the potential scale reduction factor method of Gelman et al. [8]. Posterior mean is used as the final estimate for each parameter.

3. RESULTS

In simulations, our primary aim is to demonstrate that the proposed model can detect the regulatory relations based on prior information. The proposed model is also compared to a simplified HEM [4], where the hierarchy distinguishing technical noise from the biological noise has been dropped away. Note that the error hierarchy could be added analogously to both models.

The proposed model is tested with simulated data generated from a model similar to the proposed model. We consider a model that consists of 100 genes (i), 2 conditions (j) and a single TF. The Gibbs sampling is initialized with sample mean values and is run for 1000 burn-in steps after which a sample of size 2000 is collected. The potential scale reduction factor convergence diagnostic indicates that this is typically sufficient for the Gibbs sampling to converge. Parameters are set as follows: $\mu_g = \mu_d = \mu_r = \mu_a = 0$, $\sigma_\mu^2 = 100$, $\sigma_g^2 = 1$, $\sigma_d^2 = 1$, $\sigma_r^2 = 1$, $\sigma_a^2 = 1$, $\alpha = 1$, $\beta = 0.5$. For the μ_μ parameter we use the empirical sample mean of all the measurements. Data is generated from the above model (priors) except that $\mu = 0$ and $a_i = \pm 1.5$, the additive noise ϵ is sampled from the standard normal, and 10% of z_i terms are uniformly randomly set to 1, others are 0. In the first simulation we assume to have three replicates (k) and vary $\theta_i \in \{0.5, 0.55, 0.6, 0.65\}$ for those i that corresponds to the underlying regulatory mechanisms (i.e., true $z_i = 1$) and $\theta_i \in \{0.5, 0.45, 0.4, 0.35\}$ for the others (i.e., true $z_i = 0$). In the second simulation we set $\theta_i = 0.5$ for all i and vary the number of replicates $k \in \{2, 3, 5, 10\}$. Both

simulations are repeated 50 times and average results are reported. Each individual simulation with 100 genes and varying number of replicates takes only about (on the order of) minutes to run in WinBUGS and, thus, the method should be fast enough to analyze thousands of genes.

Figure 2 (a) shows how the proposed model can detect the true regulatory relations from the simulated data with varying degrees of prior information. For the receiver operating characteristic (ROC) curves the potential target genes for the TF are estimated by ranking the genes based on the absolute magnitude of $a_i \cdot z_i$ term (averaged over posterior samples). The same figure additionally compares the performance of the model to a naive approach where the regulatory relations are estimated by computing the correlations between the estimated TF expression and all the other genes' expression measurements and picking the most strongly correlating genes as potential targets for TF regulation. While the comparison is slightly artificial, it serves as demonstration about the potential of principled data fusion approaches.

Figure 2 (b) demonstrates that as the number of replicates increases the model performance increases as well. This provides further evidence about the correct functioning of the model, but on the other hand also reveals in part that it is somewhat prone to small sample sizes. Figure 2 (b) also suggests that prior information can be more valuable than having more replicates of expression measurements.

The second important difference of the proposed model to the simplified HEM is its ability to detect differential expression that is confounded by a strongly regulating TF. In the proposed model the term r_{ij} captures the changes in gene expression due other reasons than the potential regulating TF. Since the comparison model, the simplified HEM, does not take into account any direct regulation, its estimates of r_{ij} should be erroneus in the cases where there some confounding TF regulator is present. Figure 3 presents the difference between the estimates of r_{ij}s from the proposed model and the comparison model in such cases.

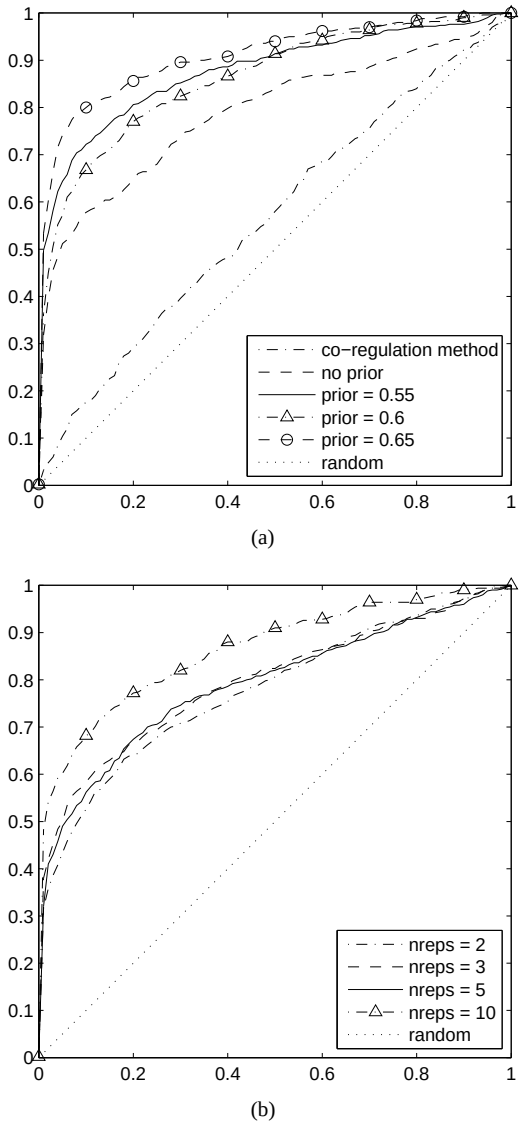


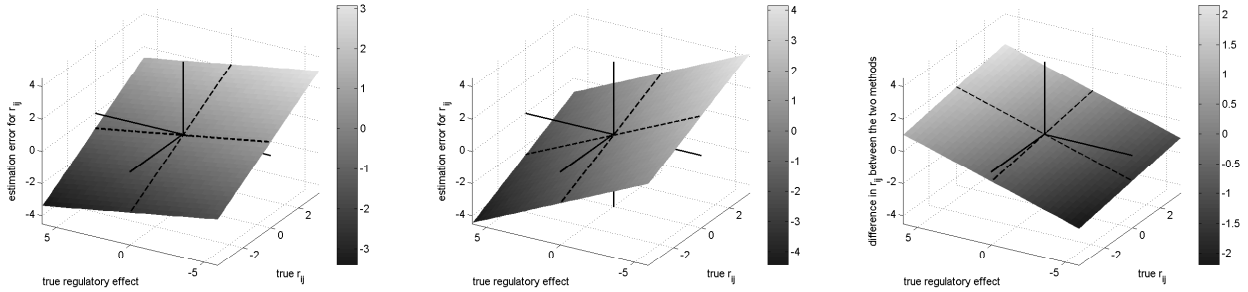
Figure 2. ROCs presenting the effect of (a) the prior strength and (b) the sample size to the model's ability to detect the true regulatory relations, in comparison to a co-expression based model.

4. DISCUSSION

We have proposed a statistical model for gene expression that can estimate separately the expression changes due to TF regulation, and the expression changes due to other reasons (unknown regulators etc.) The model is formulated in Bayesian framework and integrates the knowledge of about the potential regulators as prior data. We showed with simulated data that the model i) detects the true regulatory relations better than simple alternatives, and ii) is able to estimate the differential expression better than the comparison models in the presence of the confounding TF regulation. When there is no TF regulation present the proposed model performs equivalently to the comparison model.

The key aspect of the model is its ability to integrate prior data, such as one that describes binding probabilities of the TF proteins to the promoter regions of the genes. Since the mechanism of integration of prior information is desigend to be as simple as possible, model is versatile enough to incorporate many kinds of binding information, including for example ChIP-chip data and sequence based computationally derived binding probabilities. In particular, in the next stage, the proposed model is going to be extended to utilize a novel probabilistic model providing binding probabilities based directly on the TF motifs and promoter sequence. Since the promoter sequence is known practically for every gene for which expression can be measured, this will enable the discovery of regulatory relations for any TFs whose motifs are known, in the given experiment.

The priors we have used here represent sensible but arbitrary choices. It is clear that they have strong effects on the estimates, especially regarding the discovery of regu-



(a) The proposed model (b) The comparison model (c) Difference

Figure 3. The better ability of the proposed model to estimate differential expression in the presence of a confounding TF regulation, in comparison to the simplified HEM. Subfigures (a) and (b) show the errors in estimates from the proposed model and the comparison model for varying levels of regulation and expression changes due to other reasons (the true r_{ij}). The figures reveal the errors in r_{ij} are smaller for the proposed model than for the comparison model. Both models also make some errors in the high absolute values of true r_{ij}, which is due to selected prior centered around value zero. Subfigure (c) emphasizes how the difference between the models is largest when there is either a large positive or negative TF regulation present by showing directly the difference between the estimates of the models. Note also that the difference between the models' estimates is zero when the true regulatory effect is zero. The estimates are computed from simulated data sets including five replicates, by averaging over posterior samples and by fitting a plan for visualization purposes.

latory relations, but also with respect to other parameters. In the next stage the model will be validated more thoroughly for suitable prior distributions.

While this work focused on studying the functionality of the new model as such, the next stage will be applying the model to real gene expression data with real prior information about the binding probabilities of TFs to gene promoter regions.

5. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, (application number 213462, Finnish Programme for Centres of Excellence in Research 2006-2011).

6. REFERENCES

[1] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the USA*, vol. 95, pp. 14863–14868, 1998.

[2] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, pp. 166–176, 2003.

[3] M. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of Computational Biology*, vol. 7, pp. 819–837, 2000.

[4] H. Cho and J. K. Lee, "Bayesian hierarchical error for analysis of gene expression data," *Bioinformatics*, vol. 20, no. 13, pp. 2016–2025, 2004.

[5] H. Lähdesmäki, A. G. Rust, and I. Shmulevich, "Probabilistic inference of transcription factor binding from multiple data sources," *PLoS ONE*, vol. 3, no. 3, e1820.

[6] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon et al., "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, 2000.

[7] D. J. Lunn, A. Thomas, N. Best and D. Spiegelhalter, "WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility," *Statistics and Computing*, vol. 10, pp. 325–337, 2000.

[8] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis*, 2nd edition, Chapman & Hall/CRC, 2003.