A coupled finite mixture model for transcriptional module discovery

Han Li, Xiaodan Fan
Department of Statistics, The Chinese University of Hong Kong, HK
Corresponding author: Xiaodan Fan, e-mail: xfan@sta.cuhk.edu.hk

Abstract

Approaches to elucidate complex gene regulatory networks usually rely on the analysis of transcriptional modules (TMs). Two high-throughput technologies, gene expression microarray and Chromatin Immuno-Precipitation on Chip, often provide complementary information for discovering TMs. To efficiently integrate these two data sources, we propose a novel Bayesian model referred to as Coupled Finite Mixture Model (CFMM), which permits a separate clustering for each data source and also explicitly models their dependence. We validate our model in both a synthetic dataset and a real dataset. Our method is shown to find more consensus genes and the resulting TMs have improved biological functional coherence than those inferred by other state-of-the-art methods.

Key Words: Chip-chip data, gene expression, integrative clustering

1 Introduction

Transcriptional regulation is a crucial part of the whole regulatory mechanism in any living organism. For context-specific cellular activities, the products of genes with similar biological functions often interact with each other to form complexes, hence those genes exhibit similar expression pattern over time and space. Besides, the transcription initiation of gene expression is usually controlled through the cooperation of one or more sequence specific transcription factors that bind to the promoter sequences of their target genes. Two high-throughput technologies, gene expression microarray and Chromatin Immuno-Precipitation on Chip (Chip-chip), are capable of providing such distinct but complementary information (Eisen *et al.*, 1998; Hughes *et al.*, 2000). To unravel the underlying transcriptional modules by an optimal modeling of the two biological assays is therefore an important open problem in computation biology.

Various methods have been proposed for identifying transcription modules by integrating gene expression and transcription binding data. Most of them are clustering-based methods, including hierarchical clustering (Eisen *et al.*, 1998), K-means (Tavazoie *et al.*, 1999) and self organizing map (Tamayo *et al.*, 1999). Those methods often first cluster genes into modules based on the expression data, then attempt to find common transcription factors bound to each group of genes by some ad hoc algorithms. Genetic Regulatory Modules(GRAM) method (Bar-Joseph *et al.*, 2003) and ReMoDiscovery algorithm (Lemmens *et al.*, 2006) employ the similar technique by reversing the above two steps. However, this two-step procedure strongly relies on the assumption that co-expression and co-regulation are equivalent, which may not be valid in most cases.

Considering that gene expression and transcription binding data have their own source specific features and the inter-source correlation between them, an optimal method should

simultaneously modeling the two datasets and their dependence. Liu *et al.* (2007) developed a Bayesian hierarchical model that draws context specific clusterings and connect those clusterings through a common hyperparameter. More recently, Savage *et al.* (2010) adopted a modified hierarchical Dirichlet process mixture model to infer the context specific clusterings and also identify the consensus genes in both datasets.

Kirk *et al.* (2012) proposed a method called MDI (Multiple Dataset Integration) that assumes a finite mixture model for each dataset and captures their dependence by a parameter describing their cluster membership agreement. In the same spirit, we develop a novel Bayesian coupled finite mixture model for simultaneously modeling gene expression and Chip-chip data. Compared to MDI, our model has greater power to learn the dependence of their cluster memberships for it does not assume any specific form of the dependence. For Savage *et al.* (2010)'s method and MDI are shown to have comparable results and outperform other clustering methods in finding the transcription modules, hence, we will compare the result generated by our model with that of Savage *et al.* (2010)'s and MDI.

2 Methods

Suppose we have n experiment units, $i = 1,...,n$, and for each unit i, we have observation data y_{i1}, y_{i2} in two datasets. Denote $y_s = (y_{1s},...,y_{ns}),\ s = 1,2$, and $y = (y_1, y_2)$. We assume y_1, y_2 are generated by finite mixture models and both of them have K mixture components. For y_1, y_2 are measurements of those units at different features, they do not necessarily have the same partition structure. Nevertheless, for those features are correlated, y_1, y_2 provide disparate but complementary information for each other. Let z_{i1}, z_{i2} be the component indicators of y_{i1}, y_{i2} in their own mixture models, respectively. We would like to construct a joint distribution of z_{i1}, z_{i2} such that z_{i1}, z_{i2} are dependent and the model could learn their dependence flexibly from the data. Thus, we assume

$$p(z_{i1} = k, z_{i2} = q) = p(z_{i1} = k)p(z_{i2} = q|z_{i1} = k) = \pi_k \phi_{kq}.$$

where $0 \leq \pi_k \leq 1,\ 0 \leq \phi_{kq} \leq 1,\ \sum_{k=1}^{K} \pi_k = 1,\ \sum_{q=1}^{K} \phi_{kq} = 1$. Here we just use the basic property of joint distribution and do not presume any specific form of the dependence. Interestingly, our model includes the independent clustering model and the joint product likelihood clustering model as special cases with all $\phi_{kq} = \tilde{\pi}_q$ and all $\phi_{kq} = 1$, respectively. Denote $z_s = (z_{1s},...,z_{ns}),\ s = 1,2,\ z = (z_1, z_2),\ \phi_k = (\phi_{k1},...,\phi_{kK}),\ k = 1,...,K,$ $\phi = (\phi_1,...,\phi_K)$. Assuming given z_1 and z_2, y_1 and y_2 are independent. Here y_1 is the expression data and y_2 is the Chip-chip data.

For the expression data, suppose we have T experiment measurements for each gene and assume they follow the Gaussian distribution, that's

$$f(y_{i1}|z_{i1}, \mu, \Sigma) = f(y_{i1}|\ \mu_{z_{i1}}, \Sigma_{z_{i1}}),$$

where $\mu = (\mu_1,...,\mu_K), \Sigma = (\Sigma_1,...,\Sigma_K)$, are the component parameters.

For the Chip-chip data, assuming that we have m TFs, we would like to cluster them into the significant/nonsignificant groups. For significant TFs, they are supposed to have much larger binding probability than those nonsignificant TFs with genes in experiment, and hence they are the TFs of interest. Let $\nu_{ik} = 1$ if i-th TF is significant in the k-th cluster of Chip-chip data, otherwise $\nu_{ik} = 0$. We assume TFs in each group have the group specific binding probability, and denote them as θ_1, θ_0 for the significant/nonsignificant

group, respectively. Denote $\nu = (\nu_{11}, ..., \nu_{m1}, ..., \nu_{1K}, ..., \nu_{mK})$, $\theta = (\theta_0, \theta_1)$. Assuming those TFs are independent, we have

$$p(y_{i2}|z_{i2}, \nu, \theta) = \prod_{w=1}^{m} \theta_{\nu_{wz_{i2}}}^{y_{i2,w}} (1 - \theta_{\nu_{wz_{i2}}})^{1-y_{i2,w}}.$$

Denote $\varphi = (\pi, \phi, \mu, \Sigma, \theta)$. Adopting conjugate priors for those parameters, we obtain the full generative model as follows.

$$\begin{aligned}
&\pi \sim \text{Dir}(\alpha/K, ..., \alpha/K), \;\; \phi_k \sim \text{Dir}(\beta/K, ..., \beta/K), \\
&\mu_k \sim N(\mu_{0k}, \kappa_0 \Sigma_k), \;\; \Sigma_k \sim \text{Inverse Wishart}(\omega_0, \Lambda_0), \\
&\theta_0 \sim \text{Beta}(\gamma_{01}, \gamma_{02}), \;\; \theta_1 \sim \text{Beta}(\gamma_{11}, \gamma_{12}), \;\; \nu_{\tau k} \sim \text{Binom}(1, \zeta), \\
&z_{i1} \sim \text{Multinom}(1, \pi), \;\; z_{i2}|z_{i1} \sim \text{Multinom}(1, \phi_{z_{i1}}), \\
&f(y_{i1}, y_{i2}|z_{i1}, z_{i2}, \nu, \varphi) \sim f(y_{i1}|\mu_{z_{i1}}, \Sigma_{z_{i1}}) p(y_{i2}|z_{i2}, \nu, \theta), \\
&\qquad i = 1, ..., n, \; k = 1, ..., K, \; \tau = 1, ..., m, \; s = 1, 2.
\end{aligned}$$

In our model, since not all components are occupied in the two datasets, K just places an upper bound for the number of clusters, hence our model is not restrictive in application. We employ the Gibbs sampling algorithm to draw posterior samples for the parameters. Based on the posterior similar matrix, we extract the most likely cluster partition for each dataset using the method of Fritsch and Icstadt (2009). Besides, those genes that are consensus in both datasets are of interest, using the terminology of Savage *et al.* (2010), they are called fused genes. To find those fused genes, we first reassign the cluster labels of those genes such that the contingency table of the two clusterings satisfies that $n_{kk} \geq n_{kj}, j > k$, and $n_{11} \geq n_{22} \geq ... \geq n_{\eta\eta}$, where n_{kj} is the (k, j)-th element of the contingency table, and $\eta = \min(k_1, k_2)$ with k_1, k_2 being the number of the resulting clusters of the expression data and the Chip-chip data, respectively. Then we treat those relabeled genes having the same cluster labels in the two datasets as fused genes, or use some criteria to filter them.

3 Results

3.1 Synthetic dataset

We first generate independent $x_i \sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, $i = 1, ..., 200$. Then let $z_{i1} = 1$, if $x_{i1} \leq q_{0.25}$; $z_{i1} = 2$, if $q_{0.25} < x_{i1} \leq q_{0.5}$; $z_{i1} = 3$, if $q_{0.5} < x_{i1} \leq q_{0.75}$; $z_{i1} = 4$, if $q_{0.75} < x_{i1}$. Here q_α is the α quantile of standard normal distribution. Similarly, we set the value of z_{i2}. Here ρ controls the dependence of z_{i1}, z_{i2}. Given z_{i1}, z_{i2}, we generate y_{i1}, y_{i2} as follows.

(1) $y_{i1}|z_{i1} = 1 \sim \text{N}((0,0,0,0), I)$, $y_{i1}|z_{i1} = 2 \sim \text{N}((2,2,0,0), I)$,
$y_{i1}|z_{i1} = 3 \sim \text{N}((0,0,2,2), I)$, $y_{i1}|z_{i1} = 4 \sim \text{N}((2,2,2,2), I)$;

(2) $y_{i2}|z_{i2} = 1 \sim \text{JIB}(0.85, 0.85, 0.10, 0.10)$, $y_{i2}|z_{i2} = 2 \sim \text{JIB}(0.10, 0.10, 0.85, 0.85)$,
$y_{i2}|z_{i2} = 3 \sim \text{JIB}(0.10, 0.10, 0.10, 0.10)$, $y_{i2}|z_{i2} = 4 \sim \text{JIB}(0.85, 0.85, 0.85, 0.85)$.

Here JIB denotes the joint independent Bernoulli distributions. We compare four models: independent clustering model, joint product likelihood clustering model, MDI and CFMM. We choose different values of ρ with $\rho = 0$, 0.74, 0.93, 0.993, 1, such that the overlapping rate γ ($\gamma = \frac{\sum_{i=1}^{n} I(z_{i1} = z_{i2})}{n}$) equals 0.25, 0.50, 0.70, 0.90, 1, respectively. For each

case, we draw 5000 posterior samples and regard the first 2000 as burn in and repeat the simulation for 20 times. For measuring the cluster accuracy, we report the mean of the adjusted Rand index (ARI) (Hubert and Arabie, 1985) of different chains in each case. Besides, we also provide the sensitivity and the false discovery rate for discovering the consensus units when $\gamma \geq 0.50$.

	Gaussian data				JIB data			
	Indep.	Joint	MDI	CFMM	Indep.	Joint	MDI	CFMM
$\gamma = 0.25$	0.56	0.37	0.52	0.48	0.45	0.17	0.43	0.45
$\gamma = 0.50$	0.55	0.40	0.62	0.60	0.48	0.41	0.57	0.61
$\gamma = 0.70$	0.53	0.56	0.65	0.69	0.52	0.65	0.63	0.71
$\gamma = 0.90$	0.55	0.76	0.75	0.78	0.46	0.71	0.75	0.75
$\gamma = 1.00$	0.54	0.92	0.88	0.91	0.48	0.92	0.86	0.89

Table 1: Adjusted Rand index of the four methods at different γ values

	sensitivity		false discovery rate	
	MDI	CFMM	MDI	CFMM
$\gamma = 0.50$	0.67	0.70	0.21	0.35
$\gamma = 0.70$	0.89	0.91	0.13	0.12
$\gamma = 0.90$	0.96	0.98	0.06	0.07
$\gamma = 1.00$	0.97	0.99	0	0

Table 2: Sensitivity and false discovery rate of MDI and CFMM at different γ values

From the above two tables, we can see that CFMM has comparable performance with MDI in the ARI, sensitivity and false discovery rate measurements. Note that when $\rho = 0$ ($\gamma = 0.25$) and $\rho = 1$ ($\gamma = 1$), both CFMM and MDI have satisfactory ARI compared to the independent clustering model and the joint product likelihood clustering model, respectively, with the latter two models being the oracle model for the corresponding case. Besides, CFMM tends to report more consensus units than MDI, which results in that CFMM usually have larger sensitivity than MDI while the false discovery rates are similar except at $\gamma = 0.50$ where MDI performs better. In this simulation example, for it is sufficient to use one variable ρ to describe the agreement between these two datasets, in this sense, MDI can be regarded as the "oracle" model. The comparable results show that our model is competitive with MDI for its great flexibility in learning the dependence between the two datasets.

3.2 Yeast galactose dataset

We apply our model to a real example considered both by Savage *et al.* (2010) and Kirk *et al.* (2012). The gene expression data is taken from a subset of the expression dataset of Ideker *et al.* (2001), and this subset of genes has been extensively studied (e.g., Yeung *et al.*, 2003). It consists of 205 genes and the data was collected from 20 different perturbation experiments and each experiment contains four replicated runs. The expression patterns of those genes reflect four functional categories based on GO annotations, and this could be used to validate the cluster results. We use the average of the four replicates as the expression level and the Chip-chip data from Harbison *et al.* (2004) with significance

threshold $P = 0.001$, which provides binding information for 204 transcription factors.

For simplicity, we assume those gene expression experiments are independent, thus we use the univariate Gaussian distribution. We run the Gibbs sampling algorithm for 5000 iterations, and regard the first 2000 as burn-in. In each iteration, we consider the cluster in Chip-chip dataset as significant if at least one of its TFs is significant, and set $r_i = 1$ if gene i is assigned to one of the significant clusters. We say gene i is a fused gene if $p(r_i = 1) > 0.5$ in posterior samples. After collecting those fused genes, like the other two methods, we identify a final clustering of them by maximizing the posterior expected adjusted Rand index (Fritsch and Ickstadt, 2009).

We use the Biological Homogeneity Index (BHI; Datta and Datta, 2006) as the quality measure of the resulting TMs. Clusters with many genes share GO annotations will have high BHI score and perfect agreement will lead to a score of unity. We report the four different BHI scores by considering all categories or just the biological process(bp), cellular component(cc) and molecular function(mf) category. We identify 89 fused genes in four clusters corresponding to the four GO annotation categories with genes in the same cluster belonging to the same GO annotation category. More informative results are summarized in Figure 1 and Table 3.

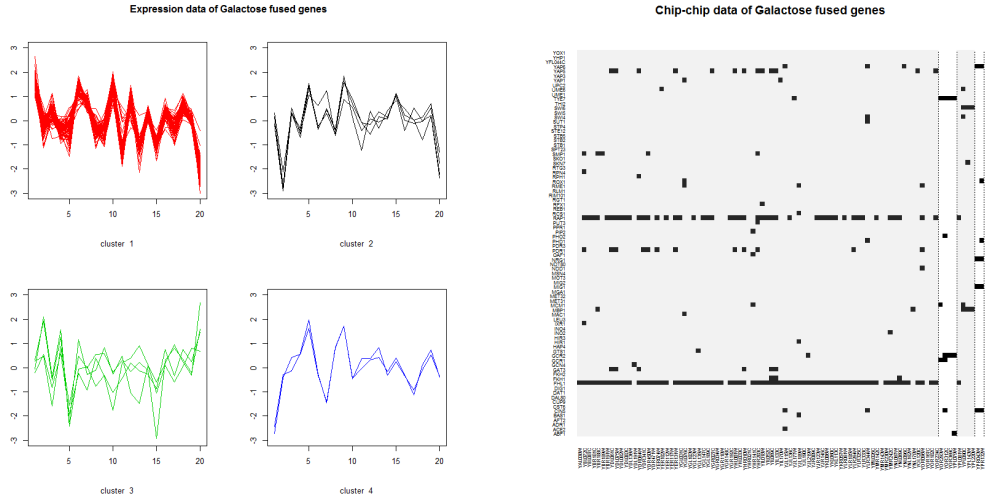

Figure 1: Expression data and Chip-chip data of Galactose fused genes

Method	BHI(all)	BHI(bp)	BHI(mf)	BHI(cc)	No. of fused genes
Savage *et al.* (2010)	0.98	0.85	0.71	0.98	72
MDI	1.00	0.89	0.77	1.00	52
CFMM	1.00	0.96	0.92	1.00	89

Table 3: Comparison of BHI scores for Savage *et al.*'s method, MDI and CFMM

From the above figure and table, we could see that our method not only reports significantly more fused genes, but also the resulting transcriptional modules have obvious cluster specific expression patterns and binding patterns, and have improved functional coherence compared with the results of the other two methods.

4 Discussion

The key innovation of our model is that it does not specify any form of the dependence between the two datasets *a prior*, but learn it from the data, and this learning ability is demonstrated in both the simulation example and the real data analysis. The results also show that by jointly modeling gene expression data and Chip-chip data, CFMM could extract more consensus genes and the resulting transcriptional modules have greater functional conformity than those inferred by the current state-of-the-art methods.

References

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. et al. (2003). Computational discovery of gene modules and regulatory networks, *Nature biotechnology* **21**(11): 1337–1342.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* **95**(25): 14863–14868.

Hubert, L. and Arabie, P. (1985). Comparing partitions, *Journal of classification* **2**(1): 193–218.

Hughes, J. D., Estep, P. W., Tavazoie, S., Church, G. M. et al. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae, *Journal of molecular biology* **296**(5): 1205–1214.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science* **292**(5518): 929–934.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets, *Bioinformatics* **28**(24): 3290–3297.

Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B. and Marchal, K. (2006). Inferring transcriptional modules from chip-chip, motif and microarray data, *Genome biology* **7**(5): R37.

Liu, X., Jessen, W. J., Sivaganesan, S., Aronow, B. J. and Medvedovic, M. (2007). Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and chip-chip data, *BMC bioinformatics* **8**(1): 283.

Savage, R. S., Ghahramani, Z., Griffin, J. E., Bernard, J. and Wild, D. L. (2010). Discovering transcriptional modules by bayesian data integration, *Bioinformatics* **26**(12): i158–i167.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proceedings of the National Academy of Sciences* **96**(6): 2907–2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999). Systematic determination of genetic network architecture, *Nature genetics* **22**(3): 281–285.