A Bayesian Nonparametric Approach to Integrative Genomics for Cancer Subgroup Discovery

Bahadir Ozdemir
Department of Computer Science
University of Maryland
College Park, MD, USA

Wael Abd-Almageed
Information Sciences Institute
University of Southern California
Arlington, VA, USA

Stephanie Roessler
Institute of Pathology
Heidelberg University Hospital
Heidelberg, Germany

Xin Wei Wang
Center for Cancer Research
National Cancer Institute
Bethesda, MD, USA

Abstract

Systematic integration of multiple omics data is a promising approach to identify cancer subgroups. In this project, miRNA and gene expression profiles were jointly analyzed for better defined molecular tumor subtypes by a Bayesian nonparametric method. Kaplan-Meier survival analysis showed that the subgroups identified by the proposed method on two hepatocellular carcinoma cohorts have different survival characteristics. Our proposal is easily extensible to other omics data types such as DNA methylation data and copy number alterations.

1 Introduction

The high tumor heterogeneity in hepatocellular carcinoma (HCC) is the major obstacle in the development of new molecularly targeted therapies. Integrated analysis of multiple genome types has become very popular to identify driving regulatory networks for certain types of cancer as the amount of available genomic data increases. Defining molecular tumor subtypes is one of the key challenges in integrative genomics. In our previous work [1], we propose a framework to identify miRNA-gene networks associated with HCC using graph mining and utilizing such networks for patient stratification in HCC using a mixture model. However, the mixture model for clustering patients dismisses the interactions between miRNAs and their target genes. In addition, a predefined number of patient clusters is required to be specified as a parameter of that model, which is quite restrictive. Here, we propose a Bayesian nonparametric model that automatically decides the number of cancer subgroups in a cohort from miRNA and gene expression profiles. Our proposed model also takes the interactions between miRNAs and their target genes into account.

Integrated analysis of omics data has gained great interest in recent years. Quantification of mRNA-miRNA interactions has been studied using a LASSO model with L1-regulatization in [2]; however, the method considers that all samples come from a homogenous cohort. Correspondence Latent Dirichlet Allocation was used for discovery of miRNA-mRNA regulatory networks in [3] and [4]. Finally, a variational Bayesian method was employed for predicting target genes of miRNAs from sequence and expression data in [5]. Unlike previous studies, our method is aiming to discover subgroups in cancer cohorts via patient clustering from microarray data. In our method, we use verified miRNA-mRNA targeting relationships and liver cancer associated genes and miRNAs.

The rest of the paper is organized as follows: Section 2 describes the proposed integrative procedure for cancer subgroup discovery. Survival analysis on the subgroups identified by the method is presented in Section 3, and Section 4 provides conclusions.

2 Our Approach

2.1 Mathematical Model

In our clustering method, we employ the first part of the iSubgraph algorithm for verifying target genes of miRNAs and selecting significant genes and miRNAs for HCC [1]. Suppose that the iSubgraph algorithm identifies G genes and M miRNAs associated with HCC, then $P_{ij} = \pm 1$ if the ith gene is a target of the jth miRNA and $P_{ij} = 0$ otherwise. $P_{ij} = +1$ indicates positively correlated targeting relationship and $P_{ij} = -1$ indicates negative correlation.

We follow the linear model in [2] for the function of miRNAs in gene regulation. Let vectors $\mathbf{u}_n = [u_{n,1}, u_{n,2}, \ldots, u_{n,G}]^T$ and $\mathbf{v}_n = [v_{n,1}, v_{n,2}, \ldots, v_{n,M}]^T$ be the logarithms of expression levels of genes and miRNAs in the nth patient, respectively. Then, the relationship between expression levels is represented by the following linear model:

$$u_{n,i} = \mu_i^u + \sum_{j=1}^{M} P_{ij}\beta_{ij}v_{n,j} + \epsilon_i \qquad \text{for } i = 1, \ldots, G \text{ and } n = 1, \ldots, N \tag{1}$$

where $\epsilon_i \sim \mathcal{N}\big(0, (\sigma_i^u)^2\big)$ is an error term, μ_i^u is the expected expression level of the ith gene in the absence of miRNAs and $\beta_{ij} \geq 0$ denotes the amount of regulation for each miRNA-gene pair. The gene and miRNA expression levels of the nth patient can be combined in a single vector for simplicity of the model as $\mathbf{x}_n = [u_{n,1}, \ldots, u_{n,G}, v_{n,1}, \ldots, v_{n,M}]^T$. Then, we define a new vector $\mathbf{y}_n$ that contains expression levels of genes and miRNAs in the absence of regulation from miRNAs as follows:

$$\mathbf{y}_n = \mathbf{B}\mathbf{x}_n, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{I}_G & -\boldsymbol{\beta} \\ \mathbf{0} & \mathbf{I}_M \end{pmatrix} \tag{2}$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix and $\boldsymbol{\beta}$ is a $G \times M$ matrix that contains regulation parameters β_{ij}. Note that $\mathbf{B}$ is invertible. Assuming that the expression level of each miRNA also follows a normal distribution, we can rewrite the model as $\mathbf{y}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $n = 1, \ldots, N$ where the covariance matrix $\boldsymbol{\Sigma}$ is diagonal. If there exist multiple subgroups, the model becomes $\mathbf{y}_n \sim \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ where $w_k = p(z_n = k)$ is a mixing proportion and z_n denotes the subgroup of the nth patient, assuming $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are shared by all subgroups.

We define a Bayesian framework for patient stratification by placing priors over w_k, β_{ij}, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$. To obtain a flexible prior, we take the infinite limit ($K \to \infty$) so that the number of exact subgroups present in a cohort becomes a variable to be determined at inference time. As a result, we have the infinite Gaussian mixture model together with a linear model for miRNA-gene interactions [6].

The infinite set of mixing proportions w_k is constructed from a stick-breaking prior (the Dirichlet process) as follows [7]:

$$w_k = w'_k \prod_{u=1}^{k-1} (1 - w'_u) \tag{3}$$

where $w'_u \sim \mathsf{Beta}(1, \alpha)$ is a proportion of the stick for all u and α is a dispersion parameter. Note that the proportions w_k sum to 1 with probability 1. Let the mean expression levels $\boldsymbol{\mu}_k$ be a random vector from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. We put an inverse-gamma prior on the variance parameter of each gene and miRNA. Let τ_i be the precision parameter for the ith element of $\mathbf{y}_n$ i.e. $\tau_i = \Sigma_{ii}^{-1}$, then we say each τ_i comes from $\mathsf{Gamma}(t_1, t_2\Sigma_{0,ii})$ where t_1 and t_2 are hyperparameters[1]. Lastly, an exponential prior is placed over each regulation parameter β_{ij} because these variables should be small since the regulation of miRNAs on genes is limited. If $P_{ij} = 0$, then β_{ij} must be fixed to zero. The regulation of a miRNA must be non-negative if $P_{ij} = +1$ and non-positive if $P_{ij} = -1$. For simplicity, we define a new variable $\beta'_{ij} = P_{ij}\beta_{ij}$ such that $\beta'_{ij} \geq 0$. All these conditions are combined in $\beta'_{ij} \sim |P_{ij}|\,\mathsf{Exponential}(\lambda) + (1 - |P_{ij}|)\,\delta_0$ where λ is a parameter that controls the regulation amount.

[1]Notation: We use shape and rate (inverse scale) parameters for Gamma distribution as $\mathsf{Gamma}(\alpha, \beta)$.

Finally, we obtain the following model (Figure 1):

$$\begin{aligned}
\mathbf{y}_n \mid z_n, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_\infty, \boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}) \\
z_n \mid w_1, \ldots, w_\infty &\sim \mathsf{Multinomial}(w_1, \ldots, w_\infty) \\
\beta'_{ij} \mid P_{ij}, \lambda &\sim |P_{ij}|\,\mathsf{Exponential}(\lambda) + (1 - |P_{ij}|)\,\delta_0 \\
\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
\tau_i \mid t_1, t_2, \boldsymbol{\Sigma}_0 &\sim \mathsf{Gamma}\big(t_1, t_2 \Sigma_{0,ii}\big) \\
w_1, \ldots, w_\infty \mid \alpha &\sim \mathsf{Stick}(1, \alpha)
\end{aligned} \tag{4}$$

where the observed data consists of expression levels $\mathbf{x}_n = \mathbf{B}^{-1}\mathbf{y}_n$. Therefore, we might replace the first line of the model with $\mathbf{x}_n \mid z_n, \boldsymbol{\beta}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_\infty, \boldsymbol{\Sigma} \sim \mathcal{N}\big(\mathbf{B}^{-1}\boldsymbol{\mu}_{z_n}, \mathbf{B}^{-1}\boldsymbol{\Sigma}\mathbf{B}^{-T}\big)$.

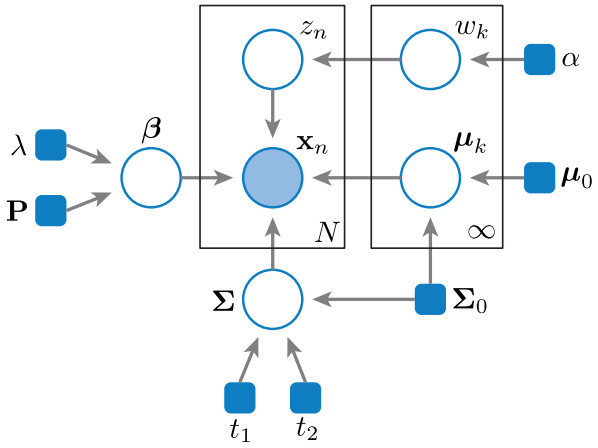

Figure 1: Graphical model for the integrative approach where circles indicate random variables, shaded circles denote observed values, and the blue square boxes are hyperparameters.

2.2 Inference by Gibbs Sampling

The posterior for z_n is not in closed form because of coupling between regulation parameters and the Dirichlet prior. Therefore, we employ a blocked Gibbs sampling procedure as [8] using a truncated Dirichlet process (TDP) in which the number of subgroups K is chosen large relative to the number patients N. The blocked Gibbs sampler iterates between the following six steps:

1. For $n \in \{1, \ldots, N\}$, independently sample z_n, the subgroup of the nth patient, from

$$p(z_n = k \,|\, \mathbf{y}_n, \mathbf{w}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \propto w_k \times \mathcal{N}(\mathbf{y}_n \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}).$$

2. For $k \in \{1, \ldots, K\}$, independently sample w'_k, the proportion of the stick, from $\mathsf{Beta}(\gamma_{k,1}, \gamma_{k,2})$ where

$$\gamma_{k,1} = 1 + \sum_{n=1}^{N} \mathbb{1}[z_n = k] \qquad \text{and} \qquad \gamma_{k,2} = \alpha + \sum_{i=k+1}^{K} \sum_{n=1}^{N} \mathbb{1}[z_n = i];$$

   then update the mixing proportion $w_k \leftarrow w'_k \prod_{i=1}^{k-1}(1 - w'_k)$.

3. For $k \in \{1, \ldots, K\}$, independently sample $\boldsymbol{\mu}_k$, mean expression levels of the kth subgroup, from $\mathcal{N}(\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k)$ where

$$\boldsymbol{\Sigma}'_k = \left(\boldsymbol{\Sigma}_0^{-1} + \sum_{n=1}^{N} \mathbb{1}[z_n = k]\, \boldsymbol{\Sigma}^{-1}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu}'_k = \boldsymbol{\Sigma}'_k \left(\boldsymbol{\Sigma}_0 \boldsymbol{\mu}_0 + \sum_{n=1}^{N} \mathbb{1}[z_n = k]\, \boldsymbol{\Sigma}^{-1} \mathbf{y}_n\right).$$

4. For $i \in \{1, \ldots, G+M\}$, independently sample τ_i, precision parameter for the ith element, from $\mathsf{Gamma}(\phi_{i,1}, \phi_{i,2})$ where

$$\phi_{i,1} = t_1 + \frac{N}{2} \qquad \text{and} \qquad \phi_{i,2} = t_2 \Sigma_{0,ii} + \frac{1}{2} \sum_{n=1}^{N} (y_{n,i} - \mu_{z_n,i})^2;$$

then update the variance $\Sigma_{ii} \leftarrow \tau_i^{-1}$.

5. For $i \in \{1, \ldots, G\}$ and $j \in \{1, \ldots, M\}$, independently sample β_{ij}, regulation parameter of the jth miRNA on the ith gene, as follows:
   (a) If $P_{ij} = 0$, do nothing (keep $\beta_{ij} = 0$),
   (b) If $P_{ij} = \pm 1$, sample β'_{ij} from a truncated normal distribution $\mathcal{N}(m_{ij}, s_{ij}^2)\mathbb{1}[\beta'_{ij} \geq 0]$ where

$$m_{ij} = \left(\sum_{n=1}^{N} \tau_i v_{n,j}^2 \right)^{-1} \left(\sum_{n=1}^{N} \left(\mu^u_{z_n,i} + \sum_{j' \neq j} \beta_{ij'} v_{n,k} - u_{n,i} \right) \tau_i v_{n,j} - P_{ij} \lambda \right),$$

$$s_{ij} = \left(\sum_{n=1}^{N} \tau_i v_{n,j}^2 \right)^{-\frac{1}{2}};$$

   then update $\beta_{ij} \leftarrow P_{ij} \beta'_{ij}$. One can sample from a truncated normal distributionq using the inverse of the normal cumulative distribution function.

6. For $n \in \{1, \ldots, N\}$ update $\mathbf{y}_n$, the expression levels in the absence of regulation, as in (2).

3 Experimental Results

We applied our approach on the same cohorts with survival data as [1], namely the Liver Cancer Institute (LCI) cohort and the Laboratory of Experimental Carcinogenesis (LEC) cohort. After applying the first part of the iSubgraph algorithm, the microarray data of the LCI cohort has 384 genes and 49 miRNAs for 196 patients. The microarray data of the LEC cohort has only 346 genes for 113 patients. The model used for the LEC cohort does not have regulation parameters ($M = 0$). Thus, it is basically a Dirichlet process Gaussian mixture model.

The hyperparameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ were set to sample average and sample covariance matrix computed from the data, respectively. We set $t_1 = N$, $t_2 = 4$ and $\lambda = 5$. The truncation level was set to $K = 1000$. The dispersion parameter α was initialized to 1 and determined by adding a Metropolis step to the Gibbs sampler. After 100 iterations, the subgroup assignments converged to 3 clusters for the LCI cohort. The Gibbs sampler with the same setting converged to 2 clusters on the LEC cohort. We used the Kaplan-Meier analysis for survival characteristics of subgroups (Figure 2). For both cohorts, the difference between survival characteristics of the subgroups identified by our method is statistically significant by the Cox-Mantel log-rank test (p-value ≤ 0.01).

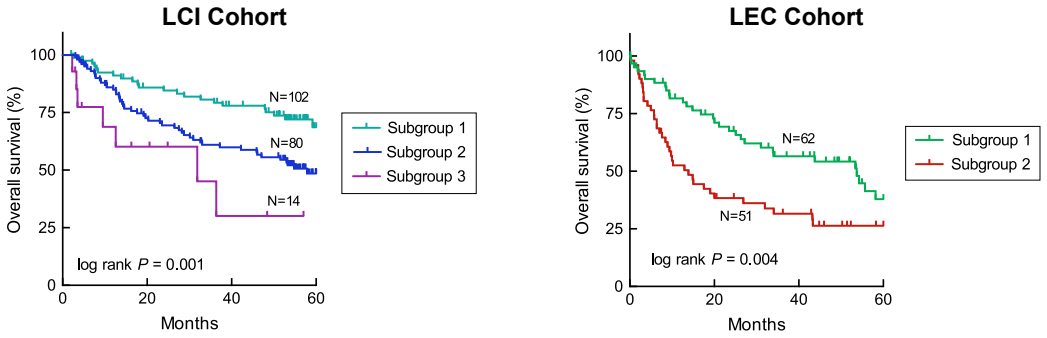

Figure 2: Kaplan-Meier plots for the subgroups identified by the method on two cohorts.

Our approach yielded similar results on the LEC cohort in comparison to the mixture model in [1]. However, our nonparametric method suggests that there exist three subgroup in the LCI cohort while

the mixture model identifies two subgroups with respect to the Bayesian information criteria. The p-value of our method on the LCI cohort for three subgroups is less than that of the mixture model (p-value $= 0.028$).

4 Conclusions

Here, we presented a Bayesian nonparametric method for integrative genomics. We used gene and miRNA expression profiles with the patterns of miRNA-gene regulations for patient stratification of two HCC cohort. Our Bayesian nonparametric approach effectively identified HCC subgroups with different survival characteristics. The advantages of this method are that it does not require a predefined number of clusters and it offers an easy extension for other genomics types such as DNA methylation or histone modification. New genomic data can be attached to the graphical model by adding new variable nodes and connecting them to other nodes which they have a relationship. In addition, the stability of class predictions might be improved by stochastic variational inference.

Acknowledgments

B. Ozdemir was supported by the UMD-NCI Partnership for Cancer Technology. This work was supported in part by grants (Z01-BC 010313) from the Intramural Research Program of the Center for Cancer Research, the National Cancer Institute.

References

[1] Bahadir Ozdemir, Wael Abd-Almageed, Stephanie Roessler, and Xin Wei Wang. iSubgraph: Integrative Genomics for Subgroup Discovery in Hepatocellular Carcinoma Using Graph Mining and Mixture Models. *PLoS ONE*, 8(11):e78624, November 2013.

[2] Ander Muniategui, Rubén Nogales-Cadenas, Miguél Vázquez, Xabier L Aranguren, Xabier Agirre, Aernout Luttun, Felipe Prosper, Alberto Pascual-Montano, and Angel Rubio. Quantification of miRNA-mRNA interactions. *PLoS ONE*, 7(2):e30766, 2012.

[3] Phaedra Agius, Yiming Ying, and Colin Campbell. Bayesian unsupervised learning with multiple data types. *Statistical applications in genetics and molecular biology*, 8(1):Article27, 2009.

[4] Bing Liu, Lin Liu, Anna Tsykin, Gregory J Goodall, Jeffrey E Green, Min Zhu, Chang Hee Kim, and Jiuyong Li. Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105–3111, December 2010.

[5] Jim C Huang, Quaid D Morris, and Brendan J Frey. Bayesian inference of MicroRNA targets from sequence and expression data. *Journal of computational biology : a journal of computational molecular cell biology*, 14(5):550–563, June 2007.

[6] Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems 12*, pages 554–560, 2000.

[7] Hemant Ishwaran and Lancelot F James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, March 2001.

[8] David M Blei and Michael I Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, March 2006.