Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2012 December ; 2012: 135–138. doi:10.1109/
GENSIPS.2012.6507747.

A Bayesian Graphical Model for Integrative Analysis of TCGA Data

Yanxun Xu,

Department of Statistics, Rice University, Houston, TX, yanxun.xu@rice.edu

Jie Zhang,

Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin

Yuan Yuan,

Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor
College of Medicine, Houston, TX

Riten Mitra,

ICES, The University of Texas at Austin, Austin, TX

Peter Müller, and

Department of Mathematics, The University of Texas at Austin, Austin, TX

Yuan Ji

CCRI, NorthShore University HealthSystem, Chicago, IL, yji@northshore.org

Abstract

We integrate three TCGA data sets including measurements on matched DNA copy numbers (C), DNA methylation (M), and mRNA expression (E) over 500+ ovarian cancer samples. The integrative analysis is based on a Bayesian graphical model treating the three types of measurements as three vertices in a network. The graph is used as a convenient way to parameterize and display the dependence structure. Edges connecting vertices infer specific types of regulatory relationships. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation. We apply the graphical model to each of the genes in the TCGA data independently and provide a comprehensive list of inferred profiles. Examples are provided based on simulated data as well.

I. Introduction

Gene expression is a critical genetic process in which DNA is transcribed to RNA. Perturbation of transcription directly affects mRNA expression and hence the subsequent protein production, leading to pathological states. Genetic variations such as copy-number variations (CNVs) and DNA methylations frequently contribute to disrupted gene

expression. CNVs result in an abnormal number of copies of DNA and thus change the gene expression level and associated phenotypes. For example, a higher copy number of CCL3L1 has been associated with lower susceptibility to HIV infection [1], and a low copy number of FCGR3B can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders [2]. DNA methylation is a biochemical modification that adds a methyl group to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. There is strong evidence that abnormal hypermethylation at the gene promoter region results in transcriptional silencing of tumor suppressor genes. Also, aberrant DNA methylation patterns have been associated with a large number of human malignancies such as cancer, lupus, and a range of birth defects [3]. Therefore, elucidating tumor-specific methylation changes will shed light on potential clinical applications in cancer diagnosis, prognosis and therapeutics [4].

Current literature mainly focuses on the pair-wise integration, between CNVs and mRNA or between methylation and mRNA. Bussey et al. [5] computed Pearson's correlation coefficients and tested the significance of correlations using false discovery rate (FDR) control. Waaijenborg et al. [6] proposed a penalized canonical correlation analysis to study genome-wide association between DNA copy number and mRNA expression. Menezes et al. [7] modeled the relationship of DNA copy number and mRNA expression by a linear model based on a modified correlation coefficient and an explorative Wilcoxon test. Choi et al. [8] described a Bayesian double-layered mixture model which directly modeled the stochastic nature of CNVs and identified abnormally expressed genes due to aberrant copy number. Etcheverry et al. [9] investigated the effect of methylation on mRNA expression in glioblastoma, and identified 13 genes that display an inverse correlation between methylation and mRNA expression using Pearson's correlation coefficient.

Since both CNVs and DNA methylation play important roles in mRNA expression, an integrated analysis that models all three platforms together is most appropriate. Denoting with C , M , and E the three platforms used to measure CNVs, methylation, and mRNA expression, we integrate data from all three platforms and present inference results as graphs that include C , M , and E as three vertices. In particular, we propose a Bayesian graphical model which imposes a probability distribution on the unknown networks and apply an autologistic prior to learn the dependence structure of three platforms through a graph. The vertices of the graph represent the platforms, and the presence or absence of edges indicates the presence or absence of conditional dependence between the platforms. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation. In this application, the use of a 3-node graphical model to represent the dependence structure of C , M and E is mainly chosen for convenience and for ease of display.

In the next Section, we give a brief overview of the ovarian cancer data to which we apply our integration analysis. In Section III, we introduce the proposed Bayesian graphical models along with MCMC simulation details. Section IV presents several simulation studies to evaluate the performance of the proposed model. In Section V, we report results based on the analysis of ovarian cancer data. We conclude with a discussion in Section VI.

II. TCGA Ovarian Cancer Data

Ovarian cancer is ranked as the fifth leading cause of death related to reproductive cancer in women. The Cancer Genome Atlas (TCGA) Research Network (<http://cancergenome.nih.gov/>) has examined more than 500 tumor samples and thousands of genes. The data is publically available online [10]. Special effort has been directed to produce matched measurements on DNA copy number (C), DNA methylation (M), and mRNA expression (E) for all the genes across the tumor samples. Taking advantage of this effort, we use the level 3 data of measurements on (C, M, E) for each gene with matched tumor samples. Specifically, let y_{itg} denote the measurement for gene g , on sample t , with platform i . Here $i = 1, 2, 3$ represents C, M , and E respectively, t indexes the $T = 534$ tumor samples, and g indexes the $N = 9283$ genes.

III. Probability Model

A. Sampling Model

We apply the proposed model for individual genes separately and thus drop the index g in subsequent discussion. For a single gene, the data is arranged in a $3 \times T$ matrix $\mathbf{Y} = [y_{it}]$, $i = 1, 2, 3$ and $t = 1, 2, \dots, T$. We assume independence of measurements y_{it} across samples. The proposed model introduces latent trinary indicators $e_{it} \in \{-1, 0, 1\}$. The indicators have an interpretation as under-, regular and over-expression of the corresponding measurement. Using e_{it} we apply the mixture model proposed by Parmigiani et al. (2002) [11] for y_{it} . In words, we assume a mixture model with uniform, normal and uniform components corresponding to under-, regular and over-expression. The model is

$$(y_{it} - \alpha_t - \mu_i) | e_{it}, \theta_{it} \sim I[e_{it} = -1]U(y_{it} | -k_{i-}, 0) + I[e_{it} = 0]N(y_{it} | 0, \sigma_i^2) + I[e_{it} = 1]U(y_{it} | 0, k_{i+}), \quad (1)$$

where $I[\cdot]$ is the indicator function, $U(A)$ denotes a uniform distribution over the set A , and $N(\cdot, \cdot)$ denotes the normal distribution. The vector $\theta_{it} = (\alpha_t, \mu_i, \sigma_i^2, k_{i-}, k_{i+})$ collects all the other parameters. For example, α_t and μ_i are the random effects of sample t and platform i . We subsequently convert the trinary variable e_{it} to a binary variable z_{it} with $p(e_{it} | z_{it} = 0) = \delta_{-1}(e_{it})$, and

$$p(e_{it} = 0 | \pi_i, z_{it} = 1) = \pi_i, p(e_{it} = 1 | \pi_i, z_{it} = 1) = 1 - \pi_i.$$

This conversion is devised to set up the following graphical model.

Denote $V = \{1, 2, 3\}$ the set of three vertices representing C, M , and E . We use a graph on these three nodes to characterize the dependence structure across the three platforms. A graph is a pair $G = \{V, S\}$ where S is a set of undirected edges $\{i, j\}$, $i, j \in V$. A graph G can be used to describe the conditional independence structure of a set of variables indexed by V , for example the binary indicators $\{z_{it}, i \in V\}$ in the case of our application. The absence of an edge $\{i, j\}$ indicates conditional independence of z_{it}, z_{jt} given the remaining variables $z_{kt}, k \neq i, k \neq j$. In the case of the three platforms the set of remaining variables reduces to just the third platform. Any joint probability model $p(z_{1t}, z_{2t}, z_{3t})$ that respects the dependence structure G can be written as (Besag, 1974 [12]):

$$p(\mathbf{z}_t|\beta, G)=p(0|\beta, G) \times \exp \left\{ \sum_{i=1}^3 \beta_i z_{it} + \sum_{\{i,j\} \in V; i < j} \beta_{ij} z_{it} z_{jt} \right\} \quad (2)$$

where $\mathbf{z}_t = (z_{1t}, z_{2t}, z_{3t})$ and $\beta = (\beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{23}, \beta_{13})$. Coefficients β_{ij} are non-zero only when the corresponding edge is included in the graph. Model (2) is known as the autologistic model.

Caragea and Kaiser [13] and Hughes et al. [14] proposed a centered parametrization of the autologistic model and argued that the centered version improves mixing of the Markov chain Monte Carlo (MCMC) posterior simulation and simplifies prior specification. The centered version is used in the form of

$$p(\mathbf{z}_t|\beta, G)=p(0|\beta, G) \exp \left\{ \sum_{i=1}^3 \beta_i z_{it} + \sum_{\{i,j\} \in V; i < j} \beta_{ij} (z_{it} - \nu_i)(z_{jt} - \nu_j) \right\}, \quad (3)$$

where $\nu_i = \exp(\beta_i) / \{1 + \exp(\beta_i)\}$.

The joint model factors as

$$p(\mathbf{Y}, \mathbf{e}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}, \beta, G) = p(\mathbf{Y}|\mathbf{e}, \boldsymbol{\theta}) p(\mathbf{e}|\mathbf{z}, \boldsymbol{\pi}) p(\mathbf{z}|\beta, G) p(\boldsymbol{\theta}) p(\beta|G) p(G) \quad (4)$$

We introduce the priors $p(\boldsymbol{\theta})p(\beta|G)p(G)$ next. Let $\text{Ga}(a, b)$ denote a gamma distribution with mean a/b . We assume conditionally conjugate priors

$$\mu_i \sim N(0, \tau_\mu), \frac{1}{\sigma_i^2} \sim \text{Ga}(\gamma_\sigma, \lambda_\sigma),$$

$$\frac{1}{k_{i-}} \sim \text{Ga}(\gamma_{k_{i-}}, \lambda_{k_{i-}}), \frac{1}{k_{i+}} \sim \text{Ga}(\gamma_{k_{i+}}, \lambda_{k_{i+}}),$$

$$\beta_* \sim N(0, \sigma_\beta^2), \pi_i \sim U(0, 1),$$

where β_* stands for the coefficients β_i, β_{ij} in (3). For the sample random effects α_t 's, we assume $\alpha_t \sim N(0, \tau_\alpha)$ subject to identifiability constraint $\sum_t \alpha_t = 0$. Lastly, we define a prior $p(G)$ as a uniform distribution over all possible graphs. With 3 vertices, we only need to consider up to 8 graphs. Each of the subgraphs is given a prior probability of 1/8.

B. Markov Chain Monte Carlo (MCMC) Simulations

We carry out posterior inference for model (4) using MCMC simulations. Each iteration of the MCMC scheme includes the following transition probabilities. We start by generating z_{it} from its complete conditional posterior. Following the update of $\mathbf{z}$, we generate values for $\mathbf{e}$ from complete conditional posterior $p(\mathbf{e} | \mathbf{Y}, \boldsymbol{\alpha}, \mathbf{z})$. If $z_{it} = 0$, the update is deterministic, $e_{it} =$

–1. If $z_{it} = 1$, the update requires a Bernoulli draw for $e_{it} = 0$ versus $e_{it} = 1$. The update of parameters θ is straight-forward. Resampling G and the regression coefficients β could be challenging in larger graphs, essentially because of the difficult evaluation of the normalization constant $p(0 | \beta, G)$ in (3) (see, e.g. [15]). However, here $p(G)$ is only supported over 8 possible graphs, making the evaluation of the normalization constant straightforward. Thus, resampling G and β reduces to straightforward trans-dimensional MCMC as in [16].

IV. Simulation Study

To evaluate the proposed model, we examine the performance of our model with 3 simulated data sets, each with $T = 300$ samples, one true graph and a single gene. For each simulation, a true graph G is first generated as follows. For a pair of vertices $\{i, j\}$, we include the edge with probability 0.5. For each imputed edge $\{i, j\}$, we generate values of $\beta_{ij} \sim N(\mu_1, 0.5^2)$ with $\mu_1 \sim U(-3, 3)$. We generate $\beta_i \sim N(\mu_2, 0.5^2)$ with $\mu_2 \sim U(-0.5, 0.5)$. Then, we generate $\mathbf{z}$ for $T = 300$ samples. Since $p(e_{it}|z_{it} = 0) = \delta_{-1}(e_{it})$, and $p(e_{it} = 0|\pi_i, z_{it} = 1) = \pi_i$, $p(e_{it} = 1|\pi_i, z_{it} = 1) = 1 - \pi_i$, we first generate $\pi_i \sim U(0.25, 0.75)$ and then generate $\mathbf{e}$. Furthermore, we let $\mu_i = 0$, $\sigma_i = 0.316$, $k_{i-} = 5.556$, $k_{i+} = 5.556$ for each node, and generate $\alpha_t \sim N(0, 0.1^2)$ subject to the identifiability criterion $\sum_t \alpha_t = 0$. Lastly, the hyper-parameters are $\tau_\alpha = 1$, $\tau_\mu = 1$, $\gamma_\sigma = 2$, $\lambda_\sigma = 0.1$, $\gamma_{k+} = 10$, $\lambda_{k+} = 50$, $\gamma_{k-} = 10$, $\lambda_{k-} = 50$, $\sigma_\beta^2 = 10$.

We implement our model to compute the posterior summaries for each simulated data set. The posterior estimates are obtained by MCMC posterior simulation with 5,000 iterations, of which 2,000 are burn-in. Since graph G is modeled as a random variable, we report the inference $\xi = P(G = G_0 | data)$, where G_0 is the true graph in the simulation. For the three data sets $\xi = 0.82, 0.86$, and 1, respectively. We also report parameter estimates $\bar{\beta} = E(\beta | \mathbf{Y})$ denoting the posterior mean for the autologistic coefficients.

From Figure 1, we can see that the estimated graph match the simulation truth for all three data sets. Here the estimated graph is the graph with highest posterior probability. We denote the positive and negative edges by black lines and red lines, respectively. The sign of β_{ij} has an intuitively appealing interpretation related to the effect of the j -th platform on the probability of presence of i -th platform, keeping the other platform fixed. Let $z_{-ij} = \mathbf{z} \setminus \{z_i, z_j\}$. We can show that β_{ij} is the log odds ratio of z_i and z_j through simple algebra, where $\beta_{ij} > 0$ implies that $p(z_i = 1 | z_j = 1, z_{-ij}) > p(z_i = 1 | z_j = 0, z_{-ij})$. See Figure 1 for the values of β 's.

V. Ovarian Cancer DATA Analysis

We apply our model and inference method to one gene at a time using the ovarian cancer data described in Section II, aiming to recover the unknown dependence structure among the three platforms for each gene, and display it as a three-vertices graph. We carry out inference using the described MCMC posterior simulation and ran 5,000 iterations with 2,000 burn-in. We obtain a posterior estimate $\hat{G}$ of the unknown graph with the largest posterior probability.

An Excel table is provided as supplementary materials in which we present the posterior probability of each subgraph for each gene (<https://sites.google.com/site/yanxunresearch>). Genes are listed in descending order according to $Pr(G = \hat{G} | data)$. There are 142 genes whose $Pr(G = \hat{G} | data) > 0.4$. When the cutoff is set to 0.6, there are 61 genes. For cutoff = 0.8, there are only 13 genes. From these 13 genes, we select two genes “ERLIN2” and “PIR” randomly to demonstrate the results.

Figure 2 shows smooth scatter plots of the data for the two selected genes. Figure 3 displays the estimated graph for them. From these two figures, we can see that the actual trend exhibited in the scatter plot is consistent with our model estimation. For example, there is an obvious positive correlation between mRNA expression and CNVs for ERLIN2 in Figure 2 and the posterior mean given by our model for the mRNA expression-CNVs edge in Figure 3 is 7.30, indicating a strong positive correlation between the two platforms, which corresponds well with what we observed in Figure 2. This matching pattern is also observed for other cases. Overall, our model estimation corresponds well with the association observed among the platforms.

VI. Discussion

We propose a Bayesian graphical model to describe the dependence structure of three genetic phenomena, CNVs, DNA methylation, and mRNA expression. The inferred graph gives a clear representation of the regulatory relationships involving the three genetic features. For example, the mRNA expression of gene ERLIN2 is sensitive to copy number changes but robust to DNA methylation, while the mRNA expression of gene PIR is sensitive to both copy number changes and DNA methylation. We are in the process of making a comprehensive list of these relationships using the entire TCGA data, expanding the effort to include more cancer types and more features such as microRNA and protein expression.

Acknowledgment

Peter Müller and Yuan Ji’s research is supported in part by NIH R01 CA132897.

References

1. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs R, Freedman B, Quinones M, Bamshad M, et al. The influence of CCL3L1 gene-containing segmental duplications on hiv-1/aids susceptibility. *Science’s STKE*. 2005; 307(5714):1434.
2. Aitman T, Dong R, Vyse T, Norsworthy P, Johnson M, Smith J, Mangion J, Robertson-Lowe C, Marshall A, Petretto E, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*. 2006; 439(7078):851–855. [PubMed: 16482158]
3. Robertson K. DNA methylation and human disease. *Nature Reviews Genetics*. 2005; 6(8):597–610.
4. Das P, Singal R. DNA methylation and cancer. *Journal of Clinical Oncology*. 2004; 22(22):4632–4642. [PubMed: 15542813]
5. Bussey K, Chin K, Lababidi S, Reimers M, Reinhold W, Kuo W, Gwadry F, Kouros-Mehr H, Fridlyand J, Jain A, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Molecular cancer therapeutics*. 2006; 5(4):853. [PubMed: 16648555]

6. Waaijenborg S, de Witt Hamer V, Philip C, Zwinderman A. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*. 2008; 7(3)
7. Menezes R, Boetzer M, Sieswerda M, Van Ommen G, Boer J. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC bioinformatics*. 2009; 10(1): 203. [PubMed: 19563656]
8. Choi H, Qin Z, Ghosh D. A double-layered mixture model for the joint analysis of DNA copy number and gene expression data. *Journal of computational biology: a journal of computational molecular cell biology*. 2010
9. Amandine E, Marc A, de Tayrac Marie V, Frederique G, Stephan S, Abderrahmane H, Laurent R, Philippe M, Veronique Q, Jean M. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. *BMC Genomics*. 2010; 11
10. Bell D, Berchuck A, Birrer M, Chien J, Cramer D, Dao F, Dhir R, Disaia P, Gabra H, Glenn P, et al. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011
11. Parmigiani G, Garrett E, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002; 64(4):717–736.
12. Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1974:192–236.
13. Caragea P, Kaiser M. Autologistic models with interpretable parameters. *Journal of agricultural, biological, and environmental statistics*. 2009; 14(3):281–300.
14. Hughes J, Haran M, Caragea P. Autologistic models for binary data on a lattice. *Environmetrics*. 2011
15. Mitra R, Mueller P, Liang S, Yue L, Ji Y. A bayesian graphical model for chip-seq data on histone modifications. *Journal of the American Statistical Association*. In Press.
16. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995; 82(4):711–732.

Fig. 1.

The simulation truth versus the estimated graph for three simulated data set. Edge colors black and red represent positive and negative relationships. The solid line represents that the edge exists. The red dotted lines indicate that the corresponding edges do not exist. The number next to each edge represents either the true value or the posterior mean of the autologistic coefficients β 's, 0 for the edges do not exist. The estimated graph based on posterior inference is identical to the simulation truth.

Fig. 2. Smooth scatter plots of pairwise relationship among platforms C , M and E . The upper panel is for gene “ERLIN2”, the low panel is for gene “PIR”. The red line in each smooth scatter plot is the lowess smoother. Dots correspond to the raw expression measurements from the level three TCGA data.

Fig. 3. Posterior estimated graphs for genes "ERLIN2" and "PIR". Black edges represent positive relationships and red edges represent negative relationships. The number next to each edge is the posterior mean of β_{ij} .