ROBUST CROSS-RACE GENE EXPRESSION ANALYSIS

Hsun-Hsien Chang and Marco F. Ramoni

Children's Hospital Informatics Program, Harvard Medical School, Boston, MA
Division of Health Sciences and Technology, Harvard-MIT, Boston, MA

ABSTRACT

This paper develops a Bayesian network (BN) predictor to profile cross-race gene expression data. Cross-race studies face more data variability than single-lab studies. Our design handles this problem by using the BN framework. In addition, unlike existing methods that unrealistically assume independent genes, our BN approach can capture the dependencies among genes. Existing BN algorithms in biomedicine applications quantize data, leading to information loss; we adopt linear Gaussian model to keep the data intact, so our resulting model is more reliable. The application of our BN predictor to a lung adenocarcinoma study shows high prediction accuracy, and performance evaluation demonstrates our gene signature agreeable with those reported in the literature. Our tool has a promising potential in finding disease biomarkers common to multiple races.

Index Terms— gene expression, Bayesian networks, transcriptional diagnosis, cross-race studies.

1. INTRODUCTION

Comparative analysis of gene expression levels between multiple tissue states makes transcriptional diagnosis feasible [1]. The analysis starts with identifying a signature of gene transcripts that differentially express across tissue conditions, and then constructs a tissue classifier using the signature. A decade ago, expression studies were conducted by single-lab analysis; i.e., the training data and the independent testing data were collected from the same research lab. Along with the advancement in microarray technology, gene expression profiling becomes a widely accepted technique in many molecular biology labs. As such, researchers can test the generalizablity of signatures beyond lab boundaries. Cross-lab studies seek biomarkers by the data acquired from one institute, and then test the predictive performance of the signature using the data obtained from another institute. Cross-lab expression analysis is more challenging because the data collected in different labs has more variability, induced by nonuniform experimental protocols such as RNA sample preparation and microarrary operations [2].

Cross-race studies are a new application area of gene expression profiling. The task of cross-race studies is to look for disease biomarkers common to multiple races. Besides having the same sources of data variability as cross-lab studies, cross-race data experience other variability due to distinct patient populations. Due to non-identical living environment, the genes in different races express diversely, so the expression levels of the same set of biomarkers vary. To handle the data variability arising from multiple data sources, we need a robust analysis tool, which is the goal of this paper.

Most of existing works were designed in the era of single-lab studies. Popular techniques can be categorized into data-driven and model-driven approaches. Data-driven methods, such as fold change [3], t statistic [4], or signal to noise ratio [5], rank all the genes based on the statistical measures of their expression levels. Model-driven methods describe the microarray data by probabilistic models and rank the genes based on a measure quantifying the model difference between tissue conditions [6]. The genes with measures exceeding an empirically determined threshold assemble a signature. Data-driven approaches are easily vulnerable to any data variability, so we opt for the model-driven approach to process multi-race data. Unlike current model based schemes, our design needs a more sophisticated model which is robust to cross-race data variability.

To avoid the predictive performance deteriorated by data variability, we consider two aspects in the design. First, we adopt a probabilistic model to describe the expression data and to regularize decision making. Second, existing methods assume that genes are independent, contradicting to the reality that genes interact directly or indirectly in biological processes. We propose to incorporate our classifier design with a more realistic network model capable of describing these dependencies. Among various design paradigms, we choose the Bayesian network (BN) framework, which is a probabilistic graph model, to meet our needs.

Besides handling data variability and capturing gene dependencies, our BN approach has the following features.

- In gene expression data, the phenotype is a discrete variable taking category numbers and the genes are continuous variables with expression levels ranging from zero to infinity. Existing BN based methods [7] in biomedicine applications quantizes variables to infer the optimal BN for the data, but quantization results in information loss. In contrast, we keep the data intact by adopting the *linear Gaussian model* to explore the dependencies among genes, yielding a more genuine BN model.
- Our signature search is capable of *eliminating collinearly expressed genes*. When a gene expresses collinearly with a biomarker, existing methods tend to include it in the signature. We avoid this problem by evaluating the likelihood of the gene's dependence on the phenotype or on another gene. If the gene is most likely dependent on the phenotype, it is a biomarker, and our BN model depicts it as modulated by the phenotype. For example, Figure 1 presents a BN describing a data set of six genes. Genes 1, 2, 3 are the biomarkers and modulated by the phenotype; genes 5 and 6 are not biomarkers but are collinear with gene 3, so the BN describes them as modulated by gene 3.
- Our BN based approach is *threshold free* to determine biomarkers. After computing scores of genes, existing works have to cut off the list by assigning a threshold. Unlike these methods that require subjective thresholds, our BN approach has determined the signature genes once the optimal network is

This research is supported in part by NIH/NHGRI (R01HG003354).

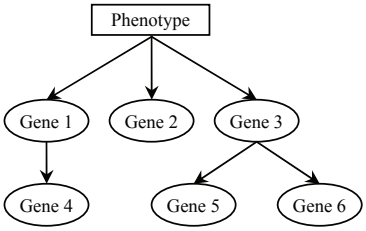

Fig. 1. Illustration of a Bayesian network.

learned from the data. The signature genes for sample classification are the genes modulated by the phenotype. Other genes not modulated by the phenotype do not play a role in classification, so they can be discarded. With reference to Figure 1, genes 1, 2, 3 assemble a signature for tissue classification; genes 4, 5, 6 can be discarded because of their irrelevance to the classification task.

2. METHODS

The BN framework for gene expression analysis consists of two steps: learn the optimal BN characterizing the given data and develop the corresponding classification scheme on testing samples. This section starts with the algorithm for learning optimal BN with linear Gaussian model, and then describes how to make prediction in our model.

2.1. Learning Bayesian Network with Linear Gaussian Model

Let $Y_1, Y_2, \cdots, Y_G$ be Gaussian random variables representing the expression levels of genes, and C be a binomial random variable characterizing two tissue conditions. We use uppercase to denote random variables and lowercase to denote their values. Given the gene expression data $\mathcal{D} = \{y_1, \cdots, y_G, c\}$, the task is to find the best BN model from a set of candidate models $\mathcal{M} = \{M_1, \cdots, M_K\}$ or, equivalently, searching for the largest posterior probability $p(M_k|\mathcal{D})$. Applying Bayes' theorem to $p(M_k|\mathcal{D})$ results in

$$p(M_k|\mathcal{D}) \propto p(M_k)p(\mathcal{D}|M_k), \tag{1}$$

where $p(M_k)$ is the prior probability of each model and $p(\mathcal{D}|M_k)$ is the *marginal* likelihood. The computation of $p(\mathcal{D}|M_k)$ is to average out θ_k from the likelihood function $p(\mathcal{D}|\theta_k)$, where Θ_k is the random vector parameterizing the distribution of $Y_1, Y_2, \cdots, Y_G, C$ conditional on M_k. We can exploit the local Markov properties encoded by the network M_k to rewrite the joint probability $p(\mathcal{D}|\theta_k)$ as

$$p(\mathcal{D}|\theta_k) = p(c|pa(c), \theta_{kc}) \prod_{g=1}^{G} p(y_g|pa(y_g), \theta_{kg}), \tag{2}$$

where $pa(y_g)$ denotes the values of the parents $Pa(Y_g)$ of Y_g, and θ_{kg} is the subset of parameters used to describe the dependence of Y_g on its parents.

In this paper, we model a gene Y_g to be dependent on either the phenotype C or another single gene Y_a, and the phenotype C is a root in the network without parents. We further can assume the J samples in the database are independent. The likelihood function becomes

$$p(\mathcal{D}|\theta_k) = \left[\prod_{j=1}^{J} p(c_j|\theta_{kc})\right] \times \left[\prod_{j=1}^{J}\prod_{g=1}^{G} p(y_{gj}|pa(y_{gj}), \theta_{kg})\right], \tag{3}$$

where the subscripts j indicate the jth sample. The first term can be estimated by the sample frequencies: $\gamma_A^{J_A}(1-\gamma_A)^{J-J_A}$, where J_A and γ_A are the number and the frequency parameter of the samples occurred in tissue condition A, respectively. The second term is computed by the linear Gaussian model [8]. When the parent of Y_g is another gene Y_a, i.e., $Pa(Y_g) = Y_a$, the conditional mean is a first order linear regression

$$\mu_g = \beta_{g0} + \beta_{g1} y_a. \tag{4}$$

When $Pa(Y_g) = C$, the conditional mean of Y_g is parameterized by c:

$$\mu_g = \beta_{g0}(c). \tag{5}$$

It follows that

$$p(y_{gj}|pa(y_{gj}), \theta_{kg}) = \left(\frac{\tau_g}{2\pi}\right)^{1/2} \exp\left(-\frac{\tau_g(y_{gj}-\mu_{gj})^2}{2}\right), \tag{6}$$

where μ_{gj} denotes the conditional mean of Y_g in sample j, and the vector θ_{kg} denotes the set of parameters $\tau_g, \beta_{g0}, \beta_{g1}$ in model M_k.

It is more convenient to adopt matrix notation to write the likelihood function in a compact form. We use the vector $\mathbf{c} = [c_1, \cdots, c_J]^T$ to denote the sample phenotypes, the vector $\mathbf{y}_g = [y_{g1}, \cdots, y_{gJ}]^T$ to stack the observations of Y_g, the vector $\beta_g = [\beta_{g0}, \beta_{g1}]^T$ to collect the regression coefficients, and the matrix

$$\mathbf{X}_g = \begin{bmatrix} 1, & pa(y_{g1}) \\ \vdots & \vdots \\ 1, & pa(y_{gJ}) \end{bmatrix} \tag{7}$$

to denote the expression values of parents of $\mathbf{y}_g$. When $Pa(Y_g) = C$, $\beta_g = [\beta_{g0}]$ and $\mathbf{X}_g = \mathbf{1}$. It follows that the second term in the likelihood function becomes

$$\prod_{g=1}^{G} \left(\frac{\tau_g}{2\pi}\right)^{J/2} \exp\left(-\frac{(\mathbf{y}_g - \mathbf{X}_g\beta_g)^T(\mathbf{y}_g - \mathbf{X}\beta_g)}{2/\tau_g}\right). \tag{8}$$

To compute the marginal likelihood, we need to learn the distributions of τ_g and β_g. The standard conjugate prior for τ_g is a Gamma distribution

$$\tau_g \sim \Gamma(\alpha_{g1}, \alpha_{g2}), \quad p(\tau_g) = \frac{1}{\alpha_{g2}^{\alpha_{g1}}\Gamma(\alpha_{g1})} \tau_g^{\alpha_{g1}-1} e^{\tau_g/\alpha_{g2}} \tag{9}$$

where $\alpha_{g1} = \frac{\nu_{g0}}{2}$ and $\alpha_{g2} = \frac{2}{\nu_{g0}\sigma_{g0}^2}$ are characterized by hyperparameters ν_{g0}, σ_{g0}^2. The marginal expectation of τ_g is

$$E(\tau_g) = \alpha_{g1}\alpha_{g2} = \frac{1}{\sigma_{g0}^2} \tag{10}$$

and

$$E(1/\tau_g) = \frac{1}{(\alpha_{g1}-1)\alpha_{g2}} = \frac{\nu_{g0}\sigma_{g0}^2}{\nu_{g0}-2} \tag{11}$$

is the prior expectation of the population variance. Because $E(1/\tau_g)$ is similar to the estimate of the variance in a sample of size ν_{g0}, σ_{g0}^2 is the prior population variance, based on ν_{g0} cases seen in the past.

Conditional on τ_g, the prior density of the parameter vector β_g is supposed to be multivariate Gaussian:

$$\beta_g|\tau_g \sim \mathcal{N}\left(\mathbf{b}_{g0}, (\tau_g \mathbf{R}_{g0})^{-1}\right) \tag{12}$$

where $\mathbf{b}_{g0} = E(\beta_g|\tau_g)$, $\mathbf{R}_{g0}$ is the identity matrix so that the regression coefficients are a priori independent, conditional on τ_g.

It can be shown that the marginal likelihood is

$$p(\{\mathbf{y}_1, \cdots, \mathbf{y}_G, \mathbf{c}\}|M_k) = \frac{1}{(2\pi)^{J/2}} \frac{|\mathbf{R}_{g0}|^{1/2}}{|\mathbf{R}_{gn}|^{1/2}} \frac{\Gamma(\nu_{gn}/2)}{\Gamma(\nu_{g0}/2)} \frac{(\nu_{g0}\sigma_{g0}^2/2)^{\nu_{g0}/2}}{(\nu_{gn}\sigma_{gn}^2/2)^{\nu_{gn}/2}} \tag{13}$$

where the parameters are specified by the following rules:

$$\alpha_{g1n} = \nu_{g0}/2 + J/2 \tag{14}$$

$$\mathbf{R}_{gn} = \mathbf{R}_{g0} + \mathbf{X}_g^T \mathbf{X}_g \tag{15}$$

$$\mathbf{b}_{gn} = \mathbf{R}_{gn}^{-1}(\mathbf{R}_{g0}\mathbf{b}_{g0} + \mathbf{X}_g^T \mathbf{y}_g) \tag{16}$$

$$\frac{1}{\alpha_{g2n}} = (-\mathbf{b}_{gn}^T \mathbf{R}_{gn} \mathbf{b}_{gn} + \mathbf{y}_g^T \mathbf{y}_g + \mathbf{b}_{g0}^T \mathbf{R}_{g0} \mathbf{b}_{g0})/2 + \frac{1}{\alpha_{g2}} \tag{17}$$

$$\nu_{gn} = \nu_{g0} + J \tag{18}$$

$$\sigma_{gn} = 2/(\nu_{gn}\alpha_{g2n}) \tag{19}$$

The Bayesian estimates of the parameters are given by the posterior expectations:

$$E(\tau_g|\mathbf{y}_g) = \alpha_{g1n}\alpha_{g2n} = 1/\sigma_{gn}^2 \tag{20}$$

$$E(\beta_g|\mathbf{y}_g) = \mathbf{b}_{gn} \tag{21}$$

$$E(1/\tau_g|\mathbf{y}_g) = \nu_{gn}\sigma_{gn}^2/(\nu_{gn} - 2) \tag{22}$$

The selection of the best BN model $\widehat{M}$ relies on the Bayes factor, BF. For arbitrary two candidate models M_k and M_h, their Bayes factor is

$$BF_{kh} = \frac{p(M_k)p(\mathcal{D}|M_k)}{p(M_h)p(\mathcal{D}|M_h)}. \tag{23}$$

If $BF_{kh} \geq 1$, we choose model $\widehat{M} = M_k$; otherwise, $\widehat{M} = M_h$. Note that when the prior distribution on the models is uniform, only the posterior odds $p(\mathcal{D}|M_k)/p(\mathcal{D}|M_h)$ contribute to the Bayes factor.

2.2. Phenotype Prediction

The phenotype prediction $\hat{c}$ of a testing sample is to find the maximum probability of the tissue class that the sample belongs to, conditional on the expression values of the sample. The formulation for the prediction is as follows:

$$\hat{c} = \arg\max_c p(c|y_1, \cdots, y_G). \tag{24}$$

The application of the Bayes' theorem to Eq. (24) gives rise to

$$\hat{c} = \arg\max_c \frac{p(y_1, \cdots, y_G|c)p(c)}{p(y_1, \cdots, y_G)} \tag{25}$$

$$= \arg\max_c p(y_1, \cdots, y_G|c)p(c), \tag{26}$$

where the second equality holds because the denominator in Eq. (25) is not a function of c. Since only genes directly dependent on the class variable C matter in the maximization, the tissue classification becomes

$$\hat{c} = \arg\max_c p(c) \prod_{g \in H} p(y_g|c), \tag{27}$$

where H denotes the set of genes that are the children of the phenotype C in the BN model.

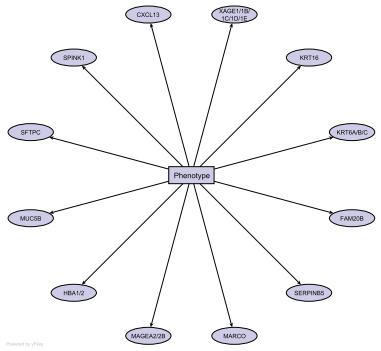

Fig. 2. The network structure learned from training data.

3. RESULTS AND DISCUSSION

We apply our method to studying molecular biomarkers of lung adenocarcinoma. The training data includes 107 subjects from the Lombardy region in Italy [9], which is publicly available on Gene Expression Omnibus (GEO) with accession number GSE10072; there are 49 controls and 58 cases. The testing data includes 63 subjects collected in Taiwan [10], which consists of 31 controls and 32 cases and whose GEO accession number is GSE7670. The gene expression experiments were carried out by Affymetrix HG-U133A, which is equipped with 22,283 probes. Probe level analysis was performed using the Robust Multi-array Algorithm (RMA). The detailed protocols of sample preparation and the demographic information of patients were described in [9, 10].

After our algorithm learns the optimal BN, we trim away the genes not modulated by the phenotype, leading to the final predictive BN model shown in Figure 2. The rectangle node is the root indicating the phenotype and the 12 elliptic nodes are signature genes. We further evaluate the prediction performance using these 12 biomarkers. The criterion for performance evaluation is the area under receiver operating characteristic (AUROC) curve. The quantity of AUROC ranges from 0 to 1; the higher the AUROC is, the better performance the predictor has. The fitted validation, i.e., predicting the training set itself, yields 100% AUROC. The prediction on the independent Taiwanese data produces 95% AUROC.

Besides the performance evaluation by AUROC, we examine the biological quality of the 12 biomarkers. Table 1 summarizes the 12 signature genes and their functions revealed in the literature. Except HBA and SPINK1, the other 10 genes are discovered to be related to lung cancer or a subtype of lung cancer, confirming good quality of our method. We briefly discuss the biomarkers in the following:

- FAM20B, MUC5B, SFTPC, and XAGE1 have been reported as biomarkers to lung adenocarcinoma.
- KRT6, KRT16, and MAGEA2 are biomarkers of squamous carcinoma; since adenocarcinoma and squamous carcinoma are both the subtypes of non-small-cell lung cancer, these 3 biomarkers explain that there is similarity between the two subtypes of lung cancer.
- CXCL13 and SERPINB5 have been known as biomarkers of lung cancer, so it is not surprised that they are predictive on adenocarcinoma.
- MARCO expresses when the lung is exposed to smoke, although it is not directly related to lung cancer. It is common that smokers have higher probability of getting lung cancer,

so MARCO is a reasonable biomarker for predicting adenocarcinoma.

- Although HBA and SPINK1 have not been reported for their association with any subtypes of lung cancer, our result suggests that it is worthwhile to study their biological function in lung cancer.

Gene Name	Function Reported in Literature
CXCL13	biomarker of lung cancer [11]
FAM20B	abundant in lung and differentially expressed in lung adenocarcinoma [12]
HBA1/HBA2	n/a
KRT6A/B/C	biomarker of lung squamous cancer [13]
KRT16	biomarker of lung squamous cancer [14]
MAGEA2/2B	biomarker of lung squamous cancer [15]
MARCO	upregulated when exposed to Lipopolysaccahrides and smoke [16]
MUC5B	biomarker of lung adenocarcinoma [17]
SERPINB5	biomarker of lung cancer [18]
SFTPC	biomarker of lung adenocarcinoma [19]
SPINK1	n/a
XAGE1/1B/1C/1D/1E	biomarker of lung adenocarcinoma [20]

Table 1. The 12-gene signature for lung adenocarcinoma diagnosis.

4. CONCLUSIONS

This paper develops a gene expression analysis algorithm in the BN framework for cross-race studies. Unlike prior works, our development adopts linear Gaussian model and considers more realistic biology that genes are dependent through their molecular interactions. The application of our BN predictor to an international lung adenocarcinoma study demonstrates how the BN method solves the real world problem. The BN predictor obtains 12 biomarkers. The prediction on an independent data set using this 12-gene signature reaches 0.95 AUROC, showing good generalizability. The biological confirmation agrees our signature with the lung cancer genes in the literature. The proposed method will have a potential to perform clinical cross-race transcriptional diagnoses.

5. REFERENCES

[1] J. Quackenbush, "Predicting the clinical status of human breast cancer by using gene expression profiles," *N. Engl. J. Med.*, vol. 354, pp. 2463–72, 2006.

[2] Members of the Toxicogenomics Research Consortium, "Standardizing global gene expression analysis between laboratories and across platforms," *Nat. Methods*, vol. 2, pp. 1–6, 2005.

[3] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Optics*, vol. 2, pp. 364–74, 1997.

[4] M. Reich, K. Ohm, M. Angelo, et al., "Genecluster 2.0: an advanced toolset for bioarray analysis," *Bioinformatics*, vol. 20, pp. 1797–8, 2004.

[5] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5116–21, 2001.

[6] P. Sebastiani, H. Xie, and M. F Ramoni, "Bayesian analysis of comparative microarray experiments by model averaging," *Bayesian Analysis*, vol. 1, pp. 707–32, 2006.

[7] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comput. Biol.*, vol. 7, pp. 601–20, 2000.

[8] F. Ferrazzi, P. Sebastiani, M. F Ramoni, and R. Bellazzi, "Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear Gaussian networks," *BMC Bioinformatics*, vol. 8, pp. e1–15, 2007.

[9] M. T. Landi, T. Dracheva, M. Rotunno, et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, pp. e1651, 2008.

[10] L.-J. Su, C.-W. Chang, Y.-C. Wu, et al., "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme," *BMC Genom.*, vol. 8, pp. 1–12, 2007.

[11] S. Singhal, D. Miller, S. Ramalingam, and S. Sun, "Gene expression profiling of non-small cell lung cancer," *Lung Cancer*, vol. 60, pp. 313–24, 2008.

[12] D. Nalbant, H. Youn, S. I. Nalbant, et al., "FAM20: an evolutionarily conserved family of secreted proteins expressed in hematopoietic cells," *BMC Genom.*, vol. 6, pp. 11, 2005.

[13] C. E. Barbieri, L. J. Tang, K. A. Brown, and J. A. Pietenpol, "Loss of p63 leads to increased cell migration and up-regulation of genes involved in invasion and metastasis," *Cancer Res.*, vol. 66, pp. 7589–97, 2006.

[14] G. D. Sgarlato, C. L. Eastman, and H. H. Sussman, "Panel of genes transcriptionally up-regulated in squamous cell carcinoma of the cervix identified by representational difference analysis, confirmed by macroarray, and validated by real-time quantitative reverse transcription-PCR," *Clin. Chem.*, vol. 51, pp. 27–34, 2005.

[15] X. Y. Zhang, Y. Hu, Y. P. Cui, et al., "Integrated genome-wide gene expression map and high-resolution analysis of aberrant chromosomal regions in squamous cell lung cancer," *FEBS Lett.*, vol. 580, pp. 2774–8, 2006.

[16] B. Sen, B. Mahadevan, and D. M. DeMarini, "Transcriptional responses to complex mixtures—A review," *Mutat. Res.*, vol. 636, pp. 144–77, 2007.

[17] M. V. Croce, A. G. Colussi, M. R. Price, and A. Segal-Eiras, "Identification and characterization of different subpopulations in a human lung adenocarcinoma cell line (a549)," *Pathol. Oncol. Res.*, vol. 5, pp. 197–204, 1999.

[18] M. Ehrich, J. K. Field, T. Liloglou, et al., C. R. Cantor, and D. van den Boom, "Cytosine methylation profiles as a molecular marker in non-small cell lung cancer," *Cancer Res.*, vol. 66, pp. 10911–8, 2006.

[19] N Nakamura, K Kobayashi, M Nakamoto, et al., "Identification of tumor markers and differentiation markers for molecular diagnosis of lung adenocarcinoma," *Oncogene*, vol. 25, pp. 4245–55, 2006.

[20] M. Shimono, A. Uenaka, Y. Noguchi, et al., "Identification of DR9-restricted XAGE antigen on lung adenocarcinoma recognized by autologous CD4 T-cells," *Int. J. Oncol.*, vol. 30, pp. 835–40, 2007.